# Marginal likelihood estimation in Spatial Statistics

Ayres, Joshua

*Award date:*
2019

# Marginal likelihood estimation in Spatial Statistics

submitted by

Joshua Ayres

for the degree of Master of Philosophy

of the

University of Bath

Department of Mathematical Sciences

June 2018

# Abstract

The marginal likelihood is a fundamental statistic for model comparison in a Bayesian framework. Spatial statistics is an area of growing importance as access to large quantities of geographical data is increasing. Bayes factors as a method for model selection is currently rarely used in the area. In this thesis several of the most prominent methods for estimating the marginal likelihood are reviewed, to see which methods would be best applied in a spatial statistics setting. It was found that the Laplace method, method of power posteriors, and steppingstone sampling gave the most promising results in initial testing. These three methods all proved to be highly accurate and precise. Laplace method is clearly the fastest of the methods, but steppingstone sampling and power posteriors should be more generalisable. However when applied to spatial models of the Matérn class specifically, due to the low dimensions of the models, the Laplace method proved to be the most effective as it was dramatically faster than the tempering methods and at least as accurate.

## Acknowledgements

I of course would like to thank my supervisors Merrilee Hurn and Gavin Shaddick for their supervision of this project and the EPSRC for the funding. I would also like to thank Adrian Hill for encouraging me when I was struggling. Thank you to my parents who have helped support me when things did not go to plan and being understanding.

# Contents

# Chapter 1

# Introduction

## 1.1 Background

Air quality and thereby air pollution is an important factor in health. There is a growing body of evidence linking disease and mortality to exposure to pollutants such as fine particulate matter ($PM_{2.5}$) and Ozone ($O_3$). There are numerous papers assessing this risk for example [Boldo et al., 2006] and [Ostro and Chestnut, 1998]. The most comprehensive overview is given by the updated air quality guidelines published by the world health organisation [WHO et al., 2006]. Although much has been done in the last ten years or so to improve air quality in more developed areas such as the USA and the European union, current levels still pose a health risk [Fann et al., 2012]. Being able to quantify problem areas will guide action that could be a significant benefit to public health.

Quantifying the levels of pollution is not a straight forward process. The information used to estimate such exposure comes from many sources including remote sensing and ground monitoring sites. Models of these data need to combine all this information from different sources, exhibiting complex spatial-temporal misalignment. The ground monitoring sites represent the most accurate measurement at a location and are taken to be the standard which other measurements are compared to [Kacenelenbogen et al., 2006], but there are many areas which are far from any of the monitoring stations. This is where data from remote sensors, typically satellites, can fill in the missing information. The measurements from the remote sensors form almost complete coverage of areas of interest but the accuracy is lower. So to combine these sources within these models and get a complete picture as accurate as possible, differences between different data sources are treated as spatially and tem-

porally dependent random effects. These then allow pollution concentrations to be predicted at any time and location with greater accuracy then using just the ground monitoring stations [Liu et al., 2005]. A Bayesian hierarchical structure allows all of these considerations to be factored into a statistical model as Bayesian hierarchical models allow a great deal of flexibility and with advances in computing power and Markov Chain Monte Carlo (MCMC) methods these models can be fit. Even with these advances fitting these models is still demanding in terms of computation time so, usually, not much is done beyond parameter estimation. Specifically, in practice model comparison through use of the marginal likelihood is not carried out for many large scale statistical models. This leads to the goal of testing methods of calculating the marginal likelihood for spatial models.

### 1.1.1 The Marginal Likelihood

The marginal likelihood, also referred to as the evidence or integrated likelihood, is a quantity that gives a standard measure of the fit of a model in a Bayesian setting. Given multiple models that give reasonable predictions, the marginal likelihood of each model will provide a way to distinguish between them. What it gives is a measure of the likelihood of the data under an assumed model. The problem is that calculating the marginal likelihood is rarely straightforward, as it is usually analytically intractable. This means that numerical methods are required to obtain estimations, and these methods may be computationally costly. This is the case with spatial models where there are usually a large number of observations in space and also possibly in time.

Here the marginal likelihood will be formally defined. Given a statistical model $m$ with prior distribution $\pi(\theta|m)$, likelihood $\pi(y|\theta, m)$ and posterior distribution $\pi(\theta|y, m)$ then $\pi(y|m)$ is the $marginal\ likelihood$ of the model. It represents the likelihood of the data given the model and is also sometimes referred to as the $evidence$ or $integrated\ likelihood$. It can be calculated by using the equality

$$\pi(y|m) = \int_{\theta} \pi(\theta|m)\pi(y|\theta, m)\,\mathrm{d}\theta. \tag{1.1.1}$$

As mentioned before this integral can be impossible to calculate directly. This leaves approximation as the only way to approach a solution generally. Approximations will be denoted $\hat{\pi}(y)$ here, where the conditioning on $m$ is omitted for notational simplicity.

Once you have the marginal likelihood then it is possible to calculate the Bayes factor which allows you to compare two models. The Bayes factor for comparing models

$m_1$ and $m_2$ is

$$BF_{12} = \frac{\pi(y|m_1)}{\pi(y|m_2)} \tag{1.1.2}$$

This is one of the most important tools for Bayesian model comparison and hypothesis testing. These Bayesian statistical, or significance, tests, now known as Bayes factors, were developed by Jeffreys and he considered them to be his chief contribution to statistics [Etz and Wagenmakers, 2015]. To cite again from
[Etz and Wagenmakers, 2015], Bayes factors are now a widely used tool finding use in disciplines such as astrophysics, forensics, psychology, economics and ecology. See [Kass and Raftery, 1995] for examples where Bayes factors can be applied.

Essentially for Bayesian models the marginal likelihood gives a clear objective comparison of models as a value that represents the likelihood of the data arising under the specified model. Bayes factors may be thought of as the Bayesian equivalent of Frequentist likelihood ratios, with the notable distinction that the models need not be nested, radically different models may be compared through their marginal likelihoods. Of course sound statistical judgement can not be replaced by simply picking the model with the highest marginal likelihood, the idea is that it provides another tool in the statistician's toolbox which can be used to aid selection of the most effective model. It is also important to consider that there may be many viable models available and it is not one single model being sought after, the marginal likelihood provides a way to weight models for model averaging, see for example [Wasserman, 2000] or [Hoeting et al., 1999].

### 1.1.2 Spatial Statistics

Data almost always has a location associated with it, the question is whether it gives information about the response variable of interest. In the case of pollution data there is a clear physical connection between levels of pollution at one site and the levels at a neighbouring site in close proximity, which suggests that distance should be taken account of if one wants to statistically model the levels of pollution over an area. Also simply because of the nature of the problem that is commonly addressed, that is we have measurements from a monitoring network and wants to infer other information about locations where there are no measurements, we have to include location in our statistical models to be able to answer important questions. Essentially we would like to know not just what the level of pollution is but also where these levels occur.

So in short, the feature that differentiates spatial data from most other statistical data is the spatial autocorrelation that it exhibits. This information needs to be incor-

porated into a statistical model if we hope to capture anything of the true nature of the data and make meaningful statements or predictions. Essentially it is about including within the correlation function of our statistical model some dependence on distance. This function will be decreasing, as simply locations far apart should exhibit lower spatial correlation than locations closer together. This is based on Tobler's intuitive first law of geography, 'everything is related to everything else, but near things are more related than distant things.' [Tobler, 1970].

# Chapter 2

# Methods for Calculating the Marginal Likelihood

## 2.1 Overview of Methods

The following sections will give a brief account of most of the methods currently available for calculating the marginal likelihood, in order of when they were developed. These methods are all possible in practice due to the development of MCMC methods allowing the simulation of draws from posterior distributions. The first method, Laplace's method, uses an asymptotic approximation and was one of the first effective methods before advances in computing hardware permitted more computationally intensive methods, such as Chib's method and annealed importance sampling, to be developed at the turn of the millennium. The nested sampling, power posteriors and steppingstone sampling methods were then designed with the intention to improve upon annealed importance sampling. Nested sampling was intended to be a more general and also more efficient method whereas power posteriors and steppingstone sampling both keep to the idea of thermodynamic integration but form the estimate of the evidence in a different manner, improving on the precision. More recently research has been done with a focus on improving the method of power posteriors by use of adaptive schemes and different quadrature rules.

## 2.2 Laplace's Method

Laplace's method of approximating integrals was applied to posterior distributions in [Tierney and Kadane, 1986]. In Laplace's method the major assumption is that the

posterior distribution is highly peaked around its maximum, so can be approximated well by a normal distribution. With a large amount of data this is often a fair assumption. However this would not be the case, for example, when the posterior has heavy tails, is skewed, or is multimodal. Nevertheless there are many situations where this is a reasonable assumption, and computationally the method is significantly cheaper than many of the alternatives [Friel and Wyse, 2012].

To derive Laplace's method in one dimension begin by defining $p(\theta|y) = \log(\pi(\theta)\pi(y|\theta))$, then the marginal likelihood can be represented by the equation

$$\pi(y) = \int \exp(p(\theta|y))\,\mathrm{d}\theta.$$

Expanding $p(\theta)$ around the posterior maximum $\hat{\theta}$ as a second order Taylor series in the above expression and using the fact that the first derivative is zero at the maximum allows a simplification of the integral. Finally to obtain the estimate it can be seen that the integrand is proportional to a Gaussian distribution, and hence the integral is known as it is the normalising constant of the Gaussian.

$$\begin{aligned}
\pi(y) &\approx \int \exp\left(p(\hat{\theta}|y) + p'(\hat{\theta}|y)(\theta-\hat{\theta}) + \frac{1}{2}p''(\hat{\theta}|y)(\theta-\hat{\theta})^2\right)\,\mathrm{d}\theta \\
&= e^{p(\hat{\theta}|y)} \int \exp\left(\frac{1}{2}p''(\hat{\theta}|y)(\theta-\hat{\theta})^2\right)\,\mathrm{d}\theta \\
&= e^{p(\hat{\theta}|y)}(2\pi)^{\frac{1}{2}}p''(\hat{\theta}|y)^{-\frac{1}{2}}.
\end{aligned}$$

This extends to multidimensional distributions with the second derivative $p''(\hat{\boldsymbol{\theta}})$ replaced by the Hessian matrix of second derivatives $H(\hat{\boldsymbol{\theta}})$. Let $k$ denote the number of dimensions of the distribution, then for the multivariate case

$$\hat{\pi}(\boldsymbol{y}) = e^{p(\hat{\boldsymbol{\theta}}|y)}(2\pi)^{\frac{k}{2}}\left|H(\hat{\boldsymbol{\theta}})^{-1}\right|^{\frac{1}{2}}. \tag{2.2.1}$$

In practice implementing (2.2.1) is straightforward as long as finding the posterior maximum is simple. When it is not simple, a numerical maximisation method may be required which does significantly increase the computational cost required to use Laplace's method if the model has a very high number of parameters. For low dimensional models it will still be significantly faster than many of the other methods, as dimension increases also simply identifying the posterior mode may become increasingly difficult. Nevertheless in practice the posterior mode will often have already been found as part of the analysis, usually being a point of interest, so the problem that

remains is to evaluate or estimate the Hessian matrix at the mode.

## 2.3 Importance Sampling Estimation

Importance sampling can be used to estimate the marginal likelihood. Taking (1.1.1), dividing and multiplying through by a distribution $\phi(\theta)$ and writing $\frac{\pi(\theta)\pi(y|\theta)}{\phi(\theta)}$ as $\omega(\theta)$ gives

$$\begin{aligned}\pi(y) &= \int_\theta \frac{\pi(\theta)\pi(y|\theta)\phi(\theta)}{\phi(\theta)}\mathrm{d}\theta = \int_\theta \omega(\theta)\phi(\theta)\mathrm{d}\theta \\ &= \mathbb{E}_\phi[\omega(\theta)] \approx \frac{1}{N}\sum_{i=1}^{N}\omega(\theta^{(i)})\end{aligned}$$

where $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}$ is a sample from the importance distribution, $\phi(\theta)$, and the $\omega(\theta^{(i)})$ are referred to as the importance weights. What distinguishes different importance sampling methods is the choice of importance distribution. There are many choices for the importance distribution, in the following subsections two examples of importance sampling estimators are shown.

### 2.3.1 Arithmetic mean estimator

The following method was presented in [Raftery et al., 1996]. Setting $\phi(\theta)$ equal to $\pi(\theta)$ yields the arithmetic mean estimator

$$\hat{\pi}(y) = \frac{1}{N}\sum_{i=1}^{N}\pi(y|\theta^{(i)}).$$

This estimator is incredibly inefficient in cases where the prior is much more diffuse than the likelihood so a very large number of points have to be generated to obtain an accurate estimate. This is because the estimate is affected mainly by just a few points that are sampled in the area of high likelihood [Xie et al., 2011]. As the prior is often diffuse relative to the likelihood it makes the arithmetic mean estimator a very poor estimator in general.

### 2.3.2 Harmonic mean estimators

Using the idea of importance sampling but instead utilising the inverse we can obtain another variety of estimator known as the harmonic mean estimator. This method was presented in [Newton and Raftery, 1994]. The general harmonic mean estimator uses a sample from the posterior and is defined as follows for a probability density $\psi(\theta)$

$$\pi(y) = \left( \mathbb{E}_{\pi(\theta|y)} \left[ \frac{\psi(\theta)}{\pi(\theta)\pi(y|\theta)} \right] \right)^{-1} \tag{2.3.1}$$

$$\approx \left( \frac{1}{N} \sum_{i=1}^{N} \frac{\psi(\theta^{(i)})}{\pi(\theta^{(i)})\pi(y|\theta^{(i)})} \right)^{-1}. \tag{2.3.2}$$

Proof of (2.3.1)

$$\begin{aligned} \mathbb{E}_{\pi(\theta|y)} \left[ \frac{\psi(\theta)}{\pi(\theta)\pi(y|\theta)} \right] &= \int_{\theta} \frac{\psi(\theta)}{\pi(\theta)\pi(y|\theta)} \frac{\pi(\theta)\pi(y|\theta)}{\pi(y)} \mathrm{d}\theta \\ &= \frac{1}{\pi(y)} \int_{\theta} \psi(\theta) \, \mathrm{d}\theta \\ &= \frac{1}{\pi(y)}. \end{aligned}$$

Setting $\psi(\theta)$ equal to $\pi(\theta)$ in (2.3.2) gives the standard harmonic mean estimator

$$\hat{\pi}(y) = \left( \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\pi(y|\theta^{(i)})} \right)^{-1} \tag{2.3.3}$$

This expression is evaluated using a sample from $\pi(\theta|y)$ which can be obtained by using a Markov chain Monte Carlo method and pointwise evaluation of the likelihood. Neither of these calculations are prohibitively expensive computationally in many cases. Unfortunately this method has drawbacks as the estimator can have unbounded variance as noted in the original paper [Newton and Raftery, 1994], and also more recently in [Neal, 2001]. This problem generally occurs in importance sampling when the importance function has fatter tails than the posterior, and specifically this is a problem in a Bayesian setting since the posterior is almost always more peaked than the prior. In addition the harmonic mean estimator is insensitive to the prior when it is known that the evidence should be very sensitive to the prior [Robert and Wraith, 2009].

## 2.4 Chib's Method

Chib's method was first introduced in [Chib, 1995] as a way to obtain the evidence from Gibbs sampling output. The equation used for estimating the evidence in Chib's method comes from rearranging Bayes' formula, evaluating at a point in the parameter space $\theta^*$ and then taking the logarithm of the equation. This gives the equality

$$\log \pi(y) = \log \pi(\theta^*) + \log \pi(y|\theta^*) - \log \pi(\theta^*|y). \tag{2.4.1}$$

For most models the prior and likelihood densities, can be evaluated exactly for given values of the parameters and observed data, the difficulty in this method arises from estimating the posterior density. The equality (2.4.1) holds for any point in the parameter space but choosing the point $\theta^*$ to be the posterior mode gives a numerically stable approximation.

The approach for estimating $\log \pi(\theta^*|y)$ will be demonstrated by considering when $\theta$ consists of three parameters $\theta_1, \theta_2, \theta_3$. Using the law of total probability

$$\begin{aligned}
\pi(\theta^*|y) &= \pi(\theta_1^*|\theta_2^*, \theta_3^*, y)\pi(\theta_2^*|\theta_3^*, y)\pi(\theta_3^*|y) \\
\log \pi(\theta^*|y) &= \log \pi(\theta_1^*|\theta_2^*, \theta_3^*, y) + \log \pi(\theta_2^*|\theta_3^*, y) + \log \pi(\theta_3^*|y).
\end{aligned}$$

Draws from these conditional distributions are used to get a Monte Carlo estimate of the unknown factors; if $\theta_1^{(1)}, ..., \theta_1^{(N)}$ and $\theta_2^{(1)}, ..., \theta_2^{(N)}$ are Gibbs sampler draws from $\theta_1$ and $\theta_2$ respectively, then

$$\begin{aligned}
\pi(\theta_3^*|y) &\approx \frac{1}{N}\sum_{i=1}^{N} \pi(\theta_3^*|\theta_1^{(i)}, \theta_2^{(i)}, y), \\
\pi(\theta_2^*|y, \theta_3^*) &\approx \frac{1}{N}\sum_{i=1}^{N} \pi(\theta_2^*|\theta_1^{(i)}, \theta_3^*, y).
\end{aligned}$$

If there are additional parameters then they are estimated in the same sequential manner.

As can be seen from above, the conditional distributions for the parameters are required, this is necessarily the case if the problem is amenable to Gibbs sampling. However, for when this is not the case the method has been extended generally to work with output from a Metropolis-Hastings sampler in much the same manner [Chib and Jeliazkov, 2001].

## 2.5 Annealed Importance Sampling

Annealed importance sampling combines the ideas of a tempering scheme, such as in [Marinari and Parisi, 1992] or [Neal, 1996], with the idea of importance sampling to estimate the evidence. The method of annealed importance sampling, for estimating normalising constants, was then developed from this and presented in [Neal, 2001].

To see how this method works, importance sampling will be looked at in a general setting where the importance function may be unnormalised. Consider importance sampling for a function $f$ using a function $g$, both operating on parameters $\theta$ where the support of $f$ is contained in the support of $g$. Writing the normalising constants of functions $f, g$ as $c_f, c_g$ respectively, to see that the ratio of their normalising constants $\frac{c_f}{c_g}$ is estimated by the mean of the importance weights observe that

$$\begin{aligned}\frac{c_f}{c_g} = \frac{1}{c_g}\int f(\theta)\,\mathrm{d}\theta &= \int \frac{f(\theta)}{g(\theta)}\frac{g(\theta)}{c_g}\,\mathrm{d}\theta \\ &= \mathbb{E}_g[\omega(\theta)] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\omega(\theta^{(i)}) \quad \text{where } \omega(\theta) = \frac{f(\theta)}{g(\theta)}. \end{aligned} \tag{2.5.1}$$

The Annealed importance sampling method requires a tempering scheme which will be denoted $t_1, t_2, ...t_M$, where $0 = t_1 < t_2 < ... < t_M = 1$ and the unnormalised distribution at $t_j$ is $\pi_j(\theta|y) \propto \pi(\theta)^{1-t_j}\pi(\theta|y)^{t_j}$. The other requirement is that there exist Markov transition kernels $T_j$, for $j = 1, 2, ..., M-1$. The Markov kernels must bridge adjacent distributions $\pi_j$ and $\pi_{j+1}$ with the requirement that each transition must leave the corresponding $\pi_j$ invariant. $\theta_1$ is a sample from $\pi_1$, $\theta_2$ is obtained from applying $T_1$ to $\theta_1$ and so on until you get $\theta_M$. Let $\tilde{T}_j$ denote the reverse transition kernel. Define distributions $f, g$ as

$$\begin{aligned} f(\theta_1, \theta_2, ..., \theta_M) &= \pi_M(\theta_M)\tilde{T}_{M-1}(\theta_M, \theta_{M-1})\tilde{T}_{M-2}(\theta_{M-1}, \theta_{M-2}) \times ... \quad \tilde{T}_1(\theta_2, \theta_1), \\ g(\theta_1, \theta_2, ..., \theta_M) &= \pi_1(\theta_1)T_1(\theta_1, \theta_2)T_2(\theta_2, \theta_3) \times ... \quad T_{M-1}(\theta_{M-1}, \theta_M). \end{aligned}$$

Therefore $f$ starts at the posterior distribution followed by a series of transitions back to the prior, $g$ starts at the prior and transitions to the posterior. Crucially, the normalising constant of $f$ is

$$c_f = \int_\theta f(\theta_1, ...\theta_M)\,\mathrm{d}\theta.$$

Substituting then applying Fubini's theorem to the above gives

$$c_f = \int_{\theta_2,\ldots,\theta_M} \pi_M(\theta_M)\tilde{T}_{M-1}(\theta_M,\theta_{M-1})\tilde{T}_{M-2}(\theta_{M-1},\theta_{M-2}) \times \ldots \\ \ldots\tilde{T}_2(\theta_3,\theta_2)\left(\int_{\theta_1}\tilde{T}_1(\theta_2,\theta_1)\mathrm{d}\theta_1\right)\mathrm{d}\theta_2,\ldots,\mathrm{d}\theta_M$$

As the transition kernels are well defined the inner integral is equal to one, and so we can apply Fubini's theorem again to gain

$$c_f = \int_{\theta_3,\ldots,\theta_M} \pi_M(\theta_M)\tilde{T}_{M-1}(\theta_M,\theta_{M-1})\tilde{T}_{M-2}(\theta_{M-1},\theta_{M-2}) \times \ldots \\ \ldots\tilde{T}_3(\theta_4,\theta_3)\left(\int_{\theta_2}\tilde{T}_2(\theta_3,\theta_2)\mathrm{d}\theta_2\right)\mathrm{d}\theta_3,\ldots,\mathrm{d}\theta_M$$

After M-1 applications this leaves us with

$$c_f = \int_{\theta_M} \pi_M(\theta_M)\,\mathrm{d}\theta_M = c_M.$$

Hence $c_f = c_M$ and by analogous argument $c_g = c_1$, where $c_j$ is the normalising constant for distribution $\pi_j$. Therefore, as can be seen from 2.5.1, using these $f$ and $g$ gives an estimate of the evidence as the average of their importance weights. To obtain an expression for this which can be used in practice f is rewritten using the fact that

$$\tilde{T}_j(\theta,\theta') = T_j(\theta',\theta)\frac{\pi_j(\theta')}{\pi_j(\theta)}.$$

Applying this to the definition of $f$,

$$f = \pi_M(\theta_M)T_{M-1}(\theta_{M-1},\theta_M)\frac{\pi_{M-1}(\theta_{M-1})}{\pi_{M-1}(\theta_M)}T_{M-2}(\theta_{M-2},\theta_{M-1})\frac{\pi_{M-2}(\theta_{M-2})}{\pi_{M-2}(\theta_{M-1})} \times \ldots \\ \ldots T_1(\theta_1,\theta_2)\frac{\pi_1(\theta_1)}{\pi_1(\theta_2)}.$$

Which after dividing through by $g$ gives the importance weight as

$$w = \frac{\pi_2(\theta_2)}{\pi_1(\theta_2)} \times \frac{\pi_3(\theta_3)}{\pi_2(\theta_3)} \times \ldots \frac{\pi_{M-1}(\theta_{M-1})}{\pi_{M-2}(\theta_{M-1})} \times \frac{\pi_M(\theta_M)}{\pi_{M-1}(\theta_M)}.$$

The benefits of annealed importance sampling are that it can handle isolated modes, although as noted in [Neal, 2001] there exists the possibility of incorrect results being obtained with no indication of a problem but this is an inherent problem of

MCMC methods in general and annealed importance sampling reduces the chance of undiscovered modes in comparison to the simpler importance sampling methods, harmonic mean or arithmetic mean, or compared to Chib's method. Annealed importance sampling works well as a general method as it will work whenever it is possible to use MCMC methods. It has lower variance than the easier to implement importance sampling methods as the importance distribution at each temperature in annealed importance sampling is close to the target distribution, which means that sampled points give more consistent weight to the overall estimate. The cost is that it is more difficult to implement and requires more computation time, around an order of magnitude higher than the computation time required to just obtain samples from the posterior. The main issue though is now that two newer methods have been developed that build on the idea from annealed importance sampling, which are steppingstone sampling and the method of power posteriors, and they appear to outperform annealed importance sampling while being similar in both complexity and computation time. These newer methods will be examined in 2.7 and 2.8.

## 2.6 Nested Sampling

Nested sampling is a method that was first proposed in [Skilling et al., 2006]. This algorithm has seen particular interest in the cosmology literature, see for example [Mukherjee et al., 2006] and has been further developed to work more efficiently for mutimodal problems in [Shaw et al., 2007] and [Feroz and Hobson, 2008], where ideas of clustering are explored. Here we consider the core of the method proposed by Skilling.

In this section the same notation will be used as in [Skilling et al., 2006] except for writing the marginal likelihood as $\pi(y)$ instead of $Z$. The marginal likelihood can be viewed as the expectation of the likelihood with respect to the prior, which can be written as

$$\pi(y) = \int_{\theta} \pi(y|\theta)\pi(\theta)\,\mathrm{d}\theta, \tag{2.6.1}$$

which we can write as

$$\int L\,\mathrm{d}X, \tag{2.6.2}$$

where $L = L(\theta)$ is the likelihood function and $\mathrm{d}X = \pi(\theta)\mathrm{d}\theta$ is the element of prior mass. The concept which gives this method this name is building up the evidence from samples within nested regions of the parameter space. The bounds that define the nested regions are where the likelihood function takes a particular value, for example

we can choose a real value $\lambda$ and we are able to define a set in the parameter space as

$$\{\, \theta \in \Omega \mid L(\theta) > \lambda \}.$$

Now, for $\lambda' > \lambda$ we have that

$$\{\, \theta \in \Omega \mid L(\theta) > \lambda' \} \subset \{\, \theta \in \Omega \mid L(\theta) > \lambda \}, \tag{2.6.3}$$

and we define a function

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta)\, \mathrm{d}\theta, \tag{2.6.4}$$

which decreases from $1$ to $0$ as $\lambda$ increases. Instead of integrating $X(\lambda)$ integrate over the inverse, which ranges from $0$ to $1$ and is much simpler to estimate numerically. From (2.6.3) we see that $X(\lambda)$ is strictly monotonic and so has an inverse. We define the inverse of $X(\lambda)$ as $q(X)$

$$\pi(y) = \int_0^1 q(X)\, \mathrm{d}X, \tag{2.6.5}$$

which can be estimated by numerical integration as

$$\sum_{i=2}^{N} (x_{i-1} - x_i) q(x_i), \tag{2.6.6}$$

where $0 < x_N < ... < x_2 < x_1 = 1$. In most cases $q(x)$ is intractable [Chopin and Robert, 2010], so $q(x_i)$ must be approximated.

## 2.7 Method of Power Posteriors

The method of power posteriors is a method that uses a tempering procedure to estimate the marginal likelihood, so called because of the relation to thermodynamic integration in the physics literature. The method of thermodynamic integration was developed in the statistical physics literature but its use for estimating the evidence was realised seperately by Lartillot and Phillipe in [Lartillot and Philippe, 2006] and by Friel and Pettitt in [Friel and Pettitt, 2008].

The idea is that a Markov Chain Monte Carlo method is used to sample a series of distributions from the prior to the posterior. This series is generated by introducing a new parameter into the posterior which is referred to as the temperature, here denoted by $t$. The choice of the values of $t$ which range from $0$ to $1$ are sometimes referred to as the tempering scheme or schedule. Estimates of the expectations of the distribu-

tions along the tempering scheme are combined to form an estimate of the marginal likelihood. In essence it changes the problem of integrating the posterior over the parameter space to integrating the power posterior with respect to the temperature, usually a large reduction in dimension.

The power posterior is defined as $p_t(\theta|y) \propto \pi(y|\theta)^t\pi(\theta)$ for $t \in [0,1]$, so it begins at the prior and transitions to the posterior by powering the likelihood component of the posterior. The power posterior has normalizing constant

$$z(y|t) = \int \pi(y|\theta)^t\pi(\theta)\mathrm{d}\theta.$$

When $t = 0$, $p_t$ is equal to the prior so $z(y|t = 0)$ is equal to the normalizing constant of the prior, i.e.. one by construction. Similarly, $z(y|t = 1)$ is equal to the normalizing constant of the posterior which means that the evidence can be written in the following manner:

$$\log \pi(y) = \log\left\{\frac{z(y|t=1)}{z(y|t=0)}\right\}. \tag{2.7.1}$$

Then the above expression can be derived by the following equality

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\log(z(y|t)) &= \frac{1}{z(y|t)}z'(y|t) \\
&= \frac{1}{z(y|t)}\frac{\mathrm{d}}{\mathrm{d}t}\int \log(\pi(y|\theta))^t\pi(\theta)\,\mathrm{d}\theta \\
&= \int \log(\pi(y|\theta))\frac{(\pi(y|\theta))^t\pi(\theta)}{z(y|t)}\,\mathrm{d}\theta \\
&= \mathrm{E}_{\pi_t}\log(\pi(y|\theta)).
\end{aligned}$$

Integrating this equation with respect to $t$ yields

$$\log\left\{\frac{z(y|t=1)}{z(y|t=0)}\right\} = \int_0^1 \mathrm{E}_{p_t}\log(\pi(y|\theta))\,\mathrm{d}t \tag{2.7.2}$$

The integral is then approximated by using a finite set of temperatures bridging between $t = 0$ and $t = 1$. Let this set be denoted $t_1, t_2, ...t_M$ where $0 = t_1 < t_2 < ... < t_M = 1$. For each $t_j$ a sample from $p_{t_j}$ is used to estimate the value of the integrand $\mathbb{E}_{p_t}[\log \pi(y|\theta)]$, which will be denoted by $E_j$. Finally, the integral is approximated using a trapezoidal rule giving

$$\log \pi(y) \approx \sum_{j=1}^{M}(t_j - t_{j-1})\left(\frac{E_{j-1} + E_j}{2}\right). \tag{2.7.3}$$

There are two kinds of error with this method, discretization error from approximating an integral with finite steps, and the Monte Carlo error. Also adjusting to each new temperature can cause underestimation or overestimation of the integral depending on whether the temperatures increase to one or decrease to zero respectively.

Adaptations to the base method have been proposed. In [Friel et al., 2014], a correction to the numerical integration method is suggested. It is shown that the variance of the samples at each point are equal to the gradient of the expected logarithmic deviance curve so that the derivatives of this curve can be obtained with very little extra work required. This allows the estimates to be improved by using a corrected trapezium rule with negligible additional computation time relative to the original algorithm. The corrected trapezium rule is, for $a, b$ in the domain of $f$ and $c \in [a, b]$,

$$\int_a^b f(x)\mathrm{d}x = (b-a)\left[\frac{f(b)+f(a)}{2}\right] - \frac{(b-a)^3}{12}f''(c). \tag{2.7.4}$$

To use this formula here, the second derivative has to be estimated, by

$$f''(c) \approx \frac{f'(b)-f'(a)}{b-a}. \tag{2.7.5}$$

So the final formula used is

$$\int_a^b f(x)\mathrm{d}x \approx (b-a)\left[\frac{f(b)+f(a)}{2}\right] - \frac{(b-a)^2}{12}\left[f'(b)-f'(a)\right]. \tag{2.7.6}$$

The other contribution in [Friel et al., 2014] is an adaptive scheme developed for selecting the temperatures to attempt to reduce the discretisation error of the estimate.

Most recently a different modification to the method of power posteriors has been proposed, in [Hug et al., 2016]. They alter the numerical integration method, using Simpson's adaptive rule instead of the trapezium rule. The reason for this is that Simpson's adaptive rule provides a relatively higher order of approximation, since the trapezoidal rule is exact for linear functions while the adaptive Simpson's rule is exact for cubic functions [Hug et al., 2016]. It requires a transformation to work effectively since as discussed previously it is beneficial to have the set of temperatures more tightly spaced towards zero rather than having uniform spacing on the whole unit interval. As others have found in this area, a power law scheduling appears to be a generally effective method for spacing the estimates [Calderhead and Girolami, 2009]. From the initial results the newer method appears to work well, although perhaps not a significant improvement over the original method in terms of the standard error of

the estimates [Hug et al., 2016], it may achieve the same level of precision with fewer function evaluations required which for very large models could save significant computation time.

## 2.8 Steppingstone sampling

Steppingstone sampling, introduced in [Xie et al., 2011], combines importance sampling and a tempering scheme. It uses the same scheme as the power posteriors method and so notation from the above section on power posteriors will be reused here. The key idea behind steppingstone sampling is that it uses importance sampling to estimate each ratio in a series bridging from the prior to the posterior distributions. The reasoning for this is that at each step the distributions are similar so there should be a reduction in variance and the overall estimate will be stabler.

Assuming the prior is normalised, the marginal likelihood is equivalent to the ratio of the normalising constants of the power posterior at $t = 1$ and $t = 0$, since the power posterior at $t = 1$ is the posterior and at $t = 0$ is the prior. Then, this ratio can be rewritten as a telescoping product of the intermediate normalising constants. That is, if $c_j$ is the normalising constant for the power posterior at temperature $t_j$ then

$$\pi(y) = \frac{c_M}{c_0} = \prod_{j=1}^{M} \frac{c_j}{c_{j-1}}.$$

Next each of these terms can be estimated using importance sampling. Specifically using $p_{j-1}(\theta|y)$ as an importance sampling function for $p_j(\theta|y)$ then

$$\frac{c_j}{c_{j-1}} \approx \frac{1}{N} \sum_{i=1}^{N} \frac{p_j(\theta_{j-1}^{(i)}|y)}{p_{j-1}(\theta_{j-1}^{(i)}|y)}.$$

Which can be simplified as

$$\frac{p_j(\theta_{j-1}^{(i)}|y)}{p_{j-1}(\theta_{j-1}^{(i)}|y)} = \frac{\pi(y|\theta)^{t_j}\pi(\theta)}{\pi(y|\theta)^{t_{j-1}}\pi(\theta)} = \pi(y|\theta_{j-1}^{(i)})^{(t_j - t_{j-1})}$$

with the sample $\theta_{j-1}^{(1)}, ..., \theta_{j-1}^{(n)}$ coming from the distribution $p_{j-1}(\theta|y)$. This means the

estimate for a single term in the product is given by

$$\frac{1}{N}\sum_{i=1}^{N}\pi(y|\theta_{j-1}^{(i)})^{(t_j-t_{j-1})}$$

These expressions are combined to give an estimate of the logarithm of the evidence. The logarithm of the evidence is almost always used in practice as it prevents underflow in the computation. Substituting the above expression into

$$\log\pi(y) = \log\frac{c_M}{c_0} = \sum_{j=1}^{M}\log\frac{c_j}{c_{j-1}}$$

gives the final expression for the estimate,

$$\log\hat{\pi}(y) = \sum_{j=1}^{M}(t_j - t_{j-1})\left[\log\sum_{i=1}^{N}\pi(y|\theta_{j-1}^{(i)}) - \log N\right].$$

## 2.9 Simulated example

To demonstrate how the methods work and to look into the effectiveness of some of the methods a simple model was considered, in particular a normal model with likelihood $y|\theta \sim N(\mu, \frac{1}{\tau})$ and priors $\mu \sim N(z, v^{-1}), \tau \sim Ga(a_0, b_0)$. While not a definitive test of the methods in practice, it gives some indicator of precision and highlights some of the potential difficulties or strengths of each method. For example the results show the lower variance of the power posterior method relative to annealed importance sampling. A box plot summarises the results, see 2-1.

The data, of size one hundred, was generated in R by the standard random normal function and can be replicated by setting the seed to ten. The hyperparameters were arbitrarily set as follows: $z = 0, v = 1, a_0 = 0.1, b_0 = 0.1$.

In this example the posterior does not resemble a known distribution so the normalising constant can not be found easily analytically. So for comparison of the methods an estimate was made of the evidence using stepping stone sampling with five thousand temperatures and ten thousand samples at each temperature. This is a higher number of temperatures then it would normally be practical to use but as the scale of this model is small it is feasible here. All the methods were then compared using this as the standard and the root mean square error scores are based on this result.

For Laplace's method, a Gibbs sampler run was used to estimate the posterior

maximiser, and then 2.2.1 gives the estimate of the evidence. In the other methods, a hundred estimates were made to assess variability; for methods that required it, five hundred and one temperatures were used spaced according to the quantiles of a Beta distribution. Explicitly, for $j = 0, 1, 2, ..., 500$ the $t_j$ were equal to the cumulative density function of a Beta$(0.3, 1)$ random variable evaluated at $j/500$, which was found to be an effective way of spacing the temperatures in [Friel and Wyse, 2012].

Table 2.1: Accuracy of methods for first example.

| Method | Mean bias | Standard deviation | RMSE |
|---|---|---|---|
| Laplace method | $2.97 \times 10^{-3}$ | - | $2.97 \times 10^{-3}$ |
| HME | $3.67$ | $4.10 \times 10^{-1}$ | $3.69$ |
| AIS | $3.26 \times 10^{-2}$ | $1.06 \times 10^{-1}$ | $3.27 \times 10^{-1}$ |
| Power posteriors | $5.31 \times 10^{-3}$ | $1.17 \times 10^{-2}$ | $1.28 \times 10^{-2}$ |
| Steppingstone sampling | $1.16 \times 10^{-4}$ | $1.07 \times 10^{-2}$ | $1.01 \times 10^{-2}$ |
| Chib's method | $-5.05 \times 10^{-4}$ | $3.86 \times 10^{-4}$ | $6.35 \times 10^{-4}$ |

### 2.9.1 Discussion for simulated example

As expected with this example the estimate given by Laplace's method is accurate and it is the fastest of the methods. This shows the strength of the method when the normal approximation is good and the number of parameters is small. Laplace's method will have limited applicability when this is not the case. In many cases where there is sufficient data the posterior will be asymptotically normal though so it will often provide accurate results.

Another easy to implement method is the standard harmonic estimator. Unfortunately the results are poor even for this simple example. As the box plot shows all of the estimates generated are above the true value. Even discounting this fact the variance would be considered unacceptably high to use in any serious model comparison.

From the tempering methods, steppingstone sampling and power posteriors methods both perform well. They are accurate at estimating the true value with low variance in their estimates. Considering annealed importance sampling, it achieves a good estimate of the evidence but seems to suffer from high variance, especially in comparison with the other two tempering methods. With all three of these methods being similar in complexity and computational cost, I would not recommend annealed importance sampling from these results.

Chib's method performs well, showing very low variance in the estimate. It shows no obvious drawbacks or weaknesses from this shallow test of its capabilities. In the

author's opinion, this method along with power posteriors and steppingstone sampling show the most promise in terms of accuracy and generality. They are all computationally costly but it is a price required for a high level of accuracy. Although saying that Laplace's method clearly performs best here, and it remains to see how well it performs in further applications with more complicated models.

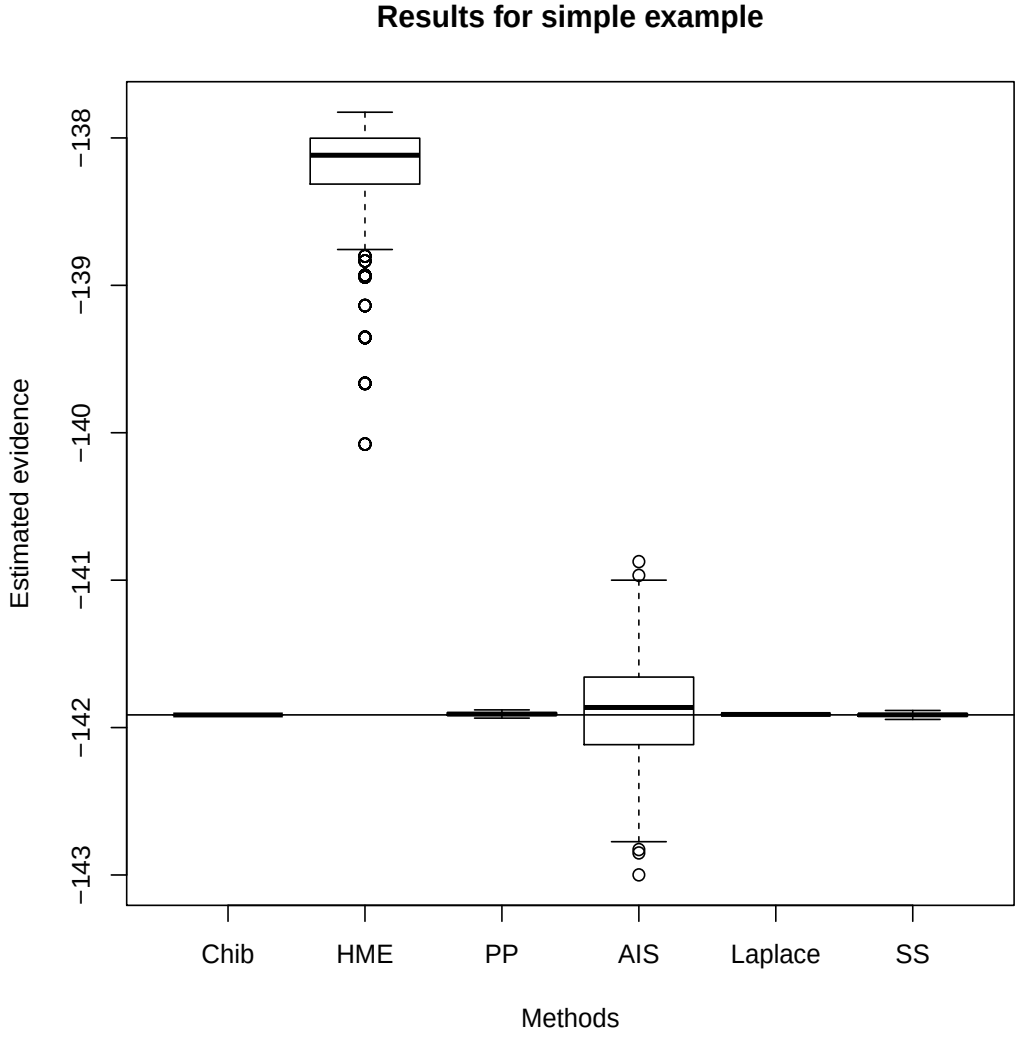



Figure 2-1: Estimates of the logarithm of the marginal likelihood for the first example. From left to right, Chib's method, Harmonic mean estimator, power posteriors, annealed importance sampling, Laplace's method, and steppingstone sampling.

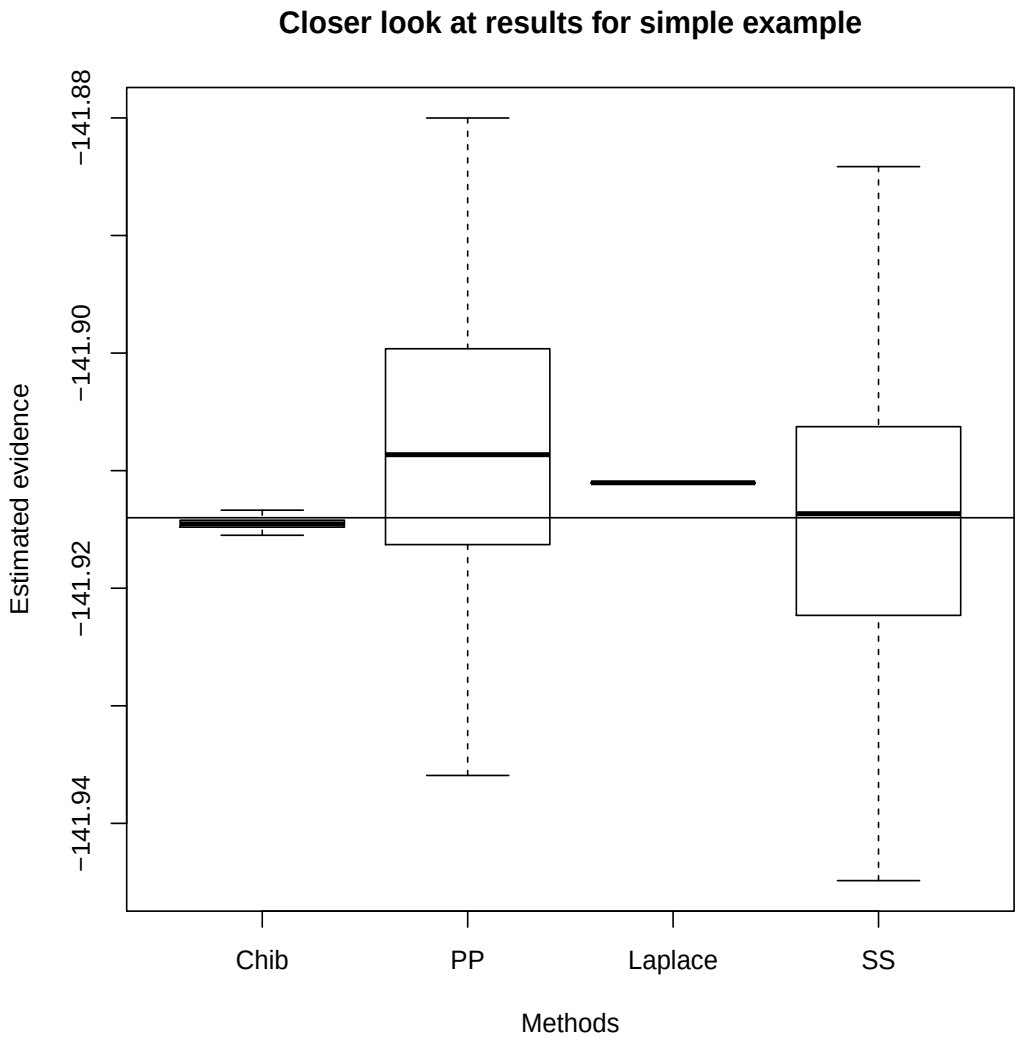



Figure 2-2: Estimates of the logarithm of the marginal likelihood for the first example for the methods with lower variance. From left to right, Chib's method, power posteriors, Laplace's method, and steppingstone sampling.

## 2.10 Second example: Pima Indians dataset

For this example we will look at real data, relating to diabetes mellitus for the Akimel O'odham population, near Phoenix, Arizona which was collected by the National Institute of Diabetes and Digestive and Kidney Diseases, and a statistical model that was used in [Smith et al., 1988]. Note that in the literature the population is referred to as Pima Indians but Akimel O'odham is the more correct nomenclature. This example was also considered in [Friel and Wyse, 2012] and in [Friel et al., 2014] so results regarding the evidence here can be compared with the ones obtained there. Calculating the marginal likelihood for this dataset will provide another benchmark of the methods for estimating the evidence. For this example only the methods that performed strongest in the previous example will be considered, to try to distinguish if these methods all perform as well on a slightly more complex problem. To be precise, Chib's method, Laplace's method, the method of power posteriors, and stepping stone sampling will be tested in this example.

The data records incidence rate of diabetes mellitus among $532$ Pima Indian women aged over twenty. The covariates in the model are number of pregnancies, plasma glucose concentration, diastolic blood pressure, body mass index, and age. The probability of incidence in each individual is modelled as a Bernoulli distribution with a Logit function relating the parameters to the probability. This was based upon the example in [Friel et al., 2014].

Unlike in the previous example, this model is not amenable to Gibb's sampling so the Metropolis Hastings algorithm was used to obtain samples from the posterior. This does not affect the implementation of Laplace's method nor the two tempering methods, but Chib's method requires a slightly different implementation, see [Chib and Jeliazkov, 2001]. For steppingstone sampling and power posterior method, fifty temperatures were used with fifty thousand samples taken at each temperature while for Chib's method a hundred thousand samples were used. For computation time this means the two tempering methods are approximately twenty five times slower than Chib's method. Laplace's method takes less than a second of computation time.

Considering the results, Chib's method appears to fare worse in this example, as can be seen from Figure 2-3, underestimating the marginal likelihood by approximately one on the logarithmic scale. The standard errors of the estimates are all low so it seems to be clearly negatively biased. This indicates a possible bias in the method for estimating the marginal likelihood from the Metropolis-Hastings output but this can not be said definitively. In the academic example in [Hug et al., 2016], Chib's method was

also found to underestimate the logarithmic marginal likelihood by a similar amount for one of the models which corroborates with the results in the second example here. The results also agree in that the estimates have a lower variance than the thermodynamic integration methods.

As can be seen from the results in Figure 2-4, Laplace's method, steppingstone sampling, and the method of power posteriors all perform well in this example with all having a small root mean square. Laplace's method seems to be the best choice here simply since it takes a lot less computation time, a matter of seconds compared to minutes for the other methods. The estimates for the method of power posteriors and steppingstone sampling perform worse on average but still to a high degree of accuracy and precision. In general when comparing the evidence of two models on the logarithmic scale, only differences of greater than a half should be considered to be meaningful [Kass and Raftery, 1995], and the standard deviation of the estimates is an order of magnitude below this threshold.

For this example the stepping stone sampling estimates and power posteriors estimates are calculated using the same Monte Carlo samples, so the results can be directly compared against each other. The stepping stone estimates were plotted against the power posterior estimates, which can be seen in Figure 2-5. The points are all clustered close to the reference $y = x$ line symmetrically indicating that the results are nearly equal for all estimates. What can be concluded from this is that in this example if the stepping stone estimate is close to the true value for a set of samples then the power posterior estimate will also be similarly accurate, and that the two estimators appear to have similar orders of standard deviation.

Table 2.2: Accuracy of methods for second example.

| Method | Mean bias | Standard deviation | Root-mean-square error |
|---|---|---|---|
| Laplace | $-1.0 \times 10^{-3}$ | - | $1.0 \times 10^{-3}$ |
| Power posteriors | $-3.5 \times 10^{-3}$ | $6.1 \times 10^{-2}$ | $4.9 \times 10^{-2}$ |
| Steppingstone sampling | $-5.3 \times 10^{-3}$ | $5.4 \times 10^{-2}$ | $4.4 \times 10^{-2}$ |
| Chib's method | $-3.80$ | $1.3$ | $3.80$ |

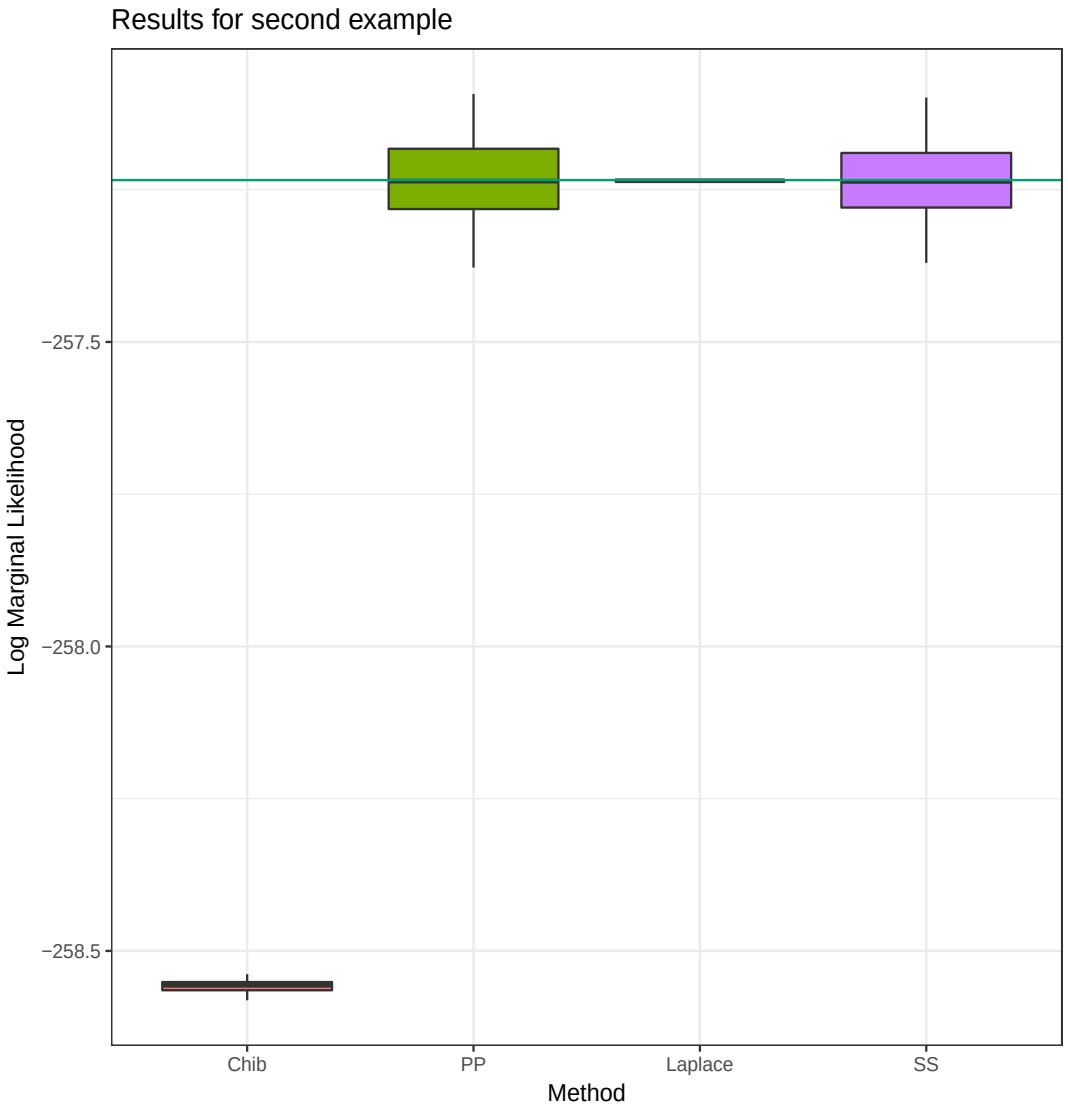



Figure 2-3: Estimated evidence for Pima Indians example with one hundred replications of each method. From left to right, Chib's method, power posteriors, Laplace's method, and steppingstone sampling. The green horizontal line represents the best approximation.

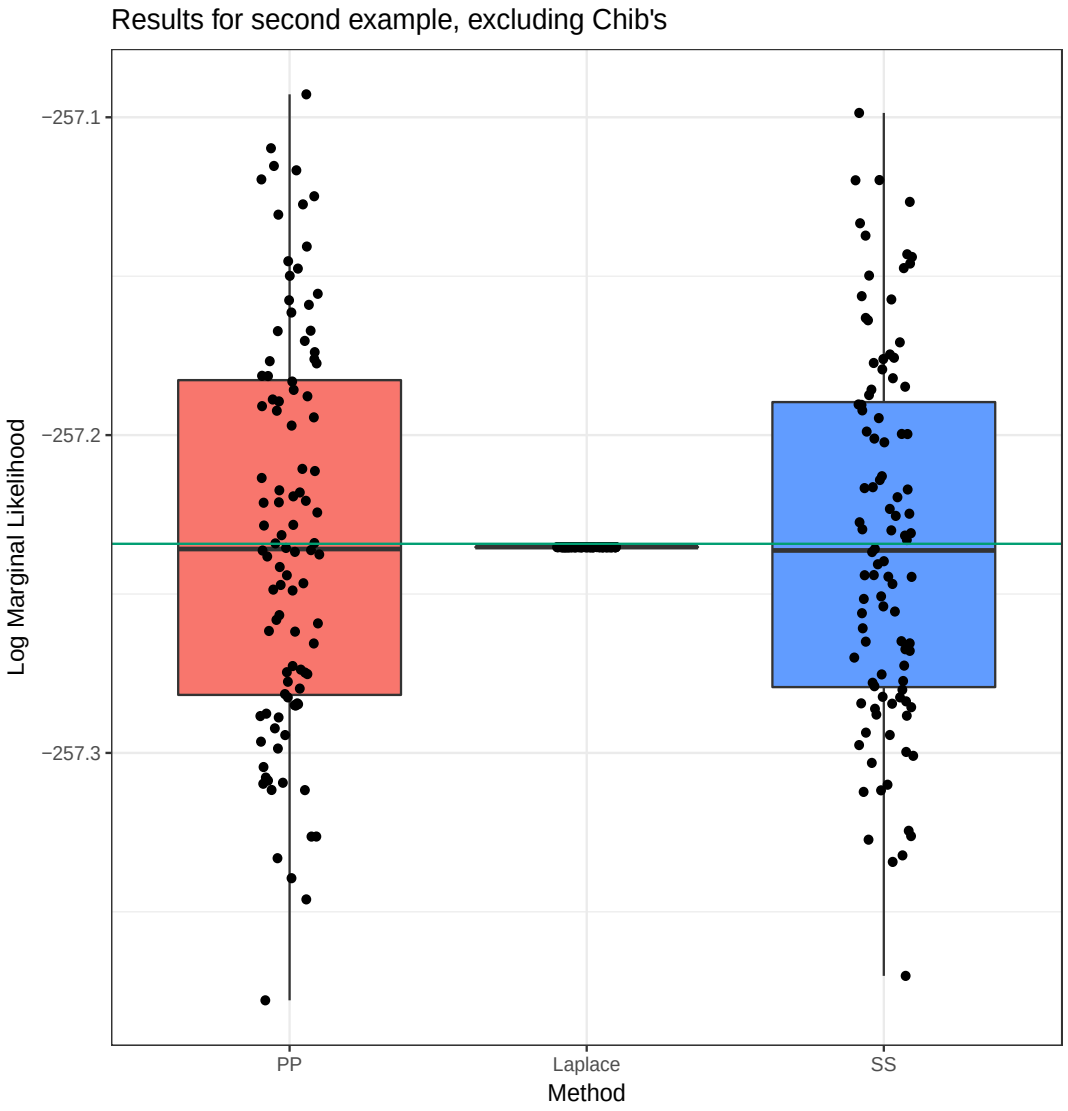



Figure 2-4: Estimated evidence for Pima Indians example excluding Chib's method. From left to right, power posteriors, Laplace's method, and steppingstone sampling. Box plot of one hundred replications with points overlaid. The green horizontal line represents the most accurate approximation.

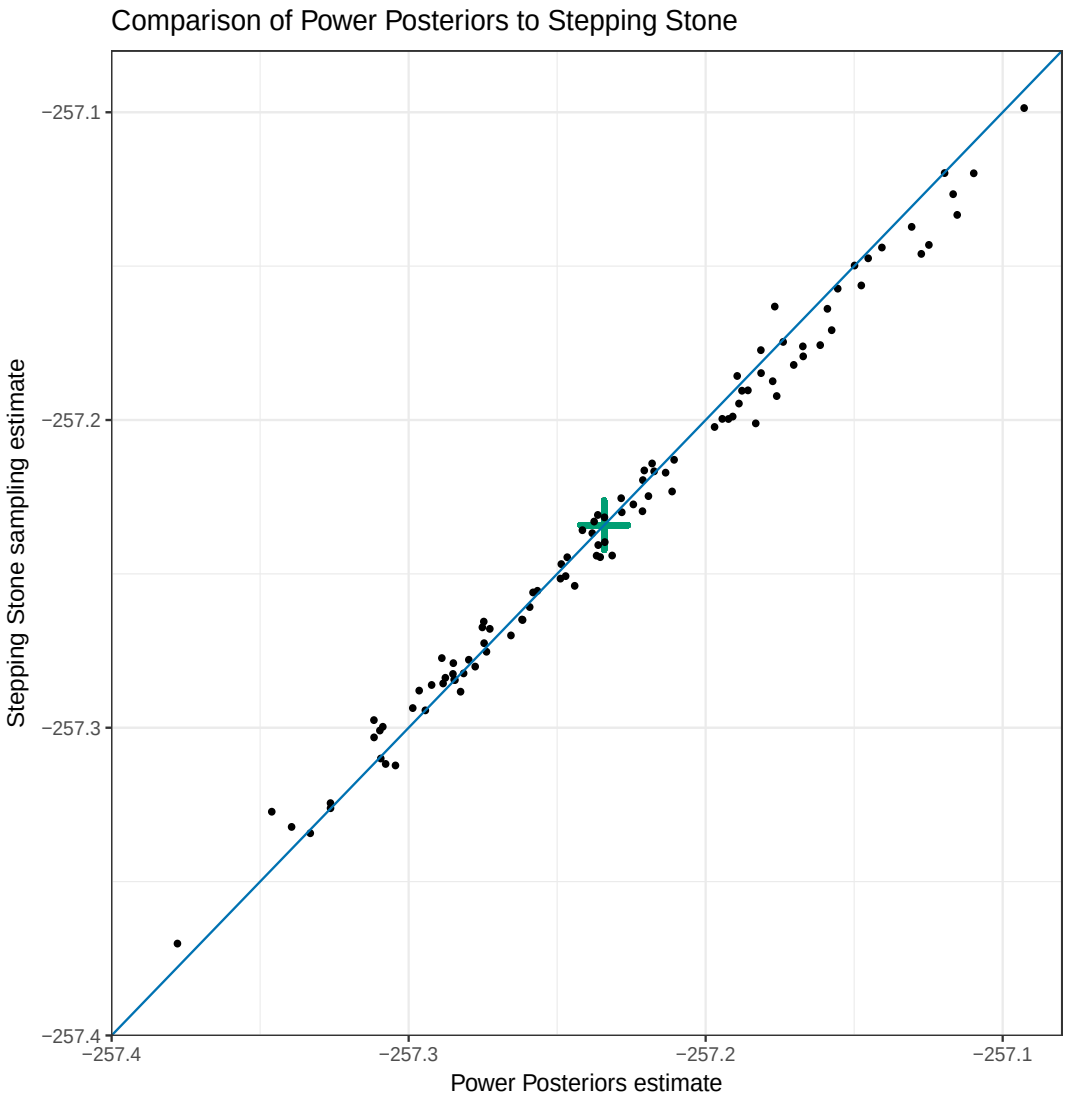



Figure 2-5: Comparison of estimates from one hundred replications, the green cross in the centre marks the target value, and the diagonal line is for reference.

### 2.10.1 A Discussion on Improving Power posteriors

In this section the results presented in the paper [Friel et al., 2014] will be considered, in particular the proposed method for improving the numerical integration component of the method of power posteriors, in the context of the current example. As shown in Section 2.7, the change to the original method involves using an improvement on the standard trapezium rule to increase the accuracy of the estimate obtained. For this improved version an estimate of the second derivative is required and, as shown in [Friel et al., 2014], estimating the second derivative is equivalent to estimating the variance. Hence the variance of the logarithmic deviance needs to be estimated at each temperature, and while this is straightforward for higher temperatures, at lower temperatures the variance tends to increase dramatically making the estimation less reliable. The reason for the increased variance at the lower temperatures is that the samples are being selected from a distribution close to the prior which is far more diffuse than the likelihood. The more diffuse the priors are relative to the likelihood, the higher the variance will be as the power posterior moves closer to $t = 0$ since at $t = 0$ the power posterior is equal to the prior. With samples coming from a distribution close to a prior which is very diffuse the logarithmic deviance of the samples will generally be very low and be highly variable. Also there will occasionally be samples from the prior close to the mode of the likelihood which will therefore have relatively high logarithmic deviance. This in turn makes the estimate of the variance unstable since these low probability points greatly influence the final estimate. It is the same issue that causes the problem that afflicts the harmonic mean estimator [Friel and Wyse, 2012]. These issues of the instability of the improved power posterior method were highlighted in [Hug et al., 2016], where it was found that the adaptive method performed worse than the regular power posteriors method in the application example. It is also important to note that the variance will be sensitive to the choice of priors, since choosing an extremely diffuse prior may cause these problems with numerical instability whereas a slightly more concentrated prior will avoid them all together.

Here we examine the effect of the correction on the estimate of the marginal likelihood via power posteriors, to see if indeed it is an improvement on the original method. Taking one estimate from the power posteriors method for this example, the correction was applied to successively more terms to see the effect it had. A graph was plotted showing the effect from only correcting one term in the numerical integration, to correcting all the terms in the summation, this is displayed in Figure 2-6. It can be seen that here the correction to the trapezium is clearly an improvement on the

base method. The estimate goes from underestimating to overestimating with the full correction, but being much closer to, the target value. From the graph it can also be inferred that the correction has smallest effect at the lower and higher temperatures where the logarithmic deviance curve is steepest and flattest respectively. These are the areas where we expect the trapezoidal approximation to the curve is closest.

For this example then the correction works as one might expect where the improved trapezium rule compensates for the clear underestimation of the standard trapezium rule and the estimate is improved upon. If however there are any problems with mixing of the Markov chains then the estimate of the variance and therefore the estimate of the second derivative will go astray turning the correction into an additional source of error. This is an inherent problem for any estimate based on samples from MCMC methods and as such is a problem for the base method as well but the problem is compounded with adding the correction term to the trapezium rule. Assuming the chains mix well the problem particular to the correction and not the base method is that the variance of the logarithmic deviance at low temperatures may simply be too high to estimate to a high enough degree of accuracy from a practical number of MCMC samples. This is perhaps the source of increased standard deviation of the estimates in the application example in [Hug et al., 2016] where they test the improved power posterior method, although this is speculative since they are additionally using adaptive rung placement which may be the cause of the larger error and not the modified trapezium rule. Nevertheless this example highlights the fact that trying to improve the original power posterior method may in fact lead to a poorer estimate. The issue is largely due to constraints on computation time, if more temperatures are used and more samples are taken at each temperature then these problems may not come to light at all since there will be enough to samples to accurately estimate the variance. But if the number of samples that can be taken are limited by such practical constraints, as they often are, then it is worth investigating whether the correction is actually correcting the estimate. The correction may be effective at the higher temperatures so simply checking the variance and selecting a cutoff may prove to give the biggest benefit if there are issues with large variance at the lower temperatures. That is to say, use the correction on all terms down to a certain value of $t$ where the variance appears too high to estimate given the number of samples.

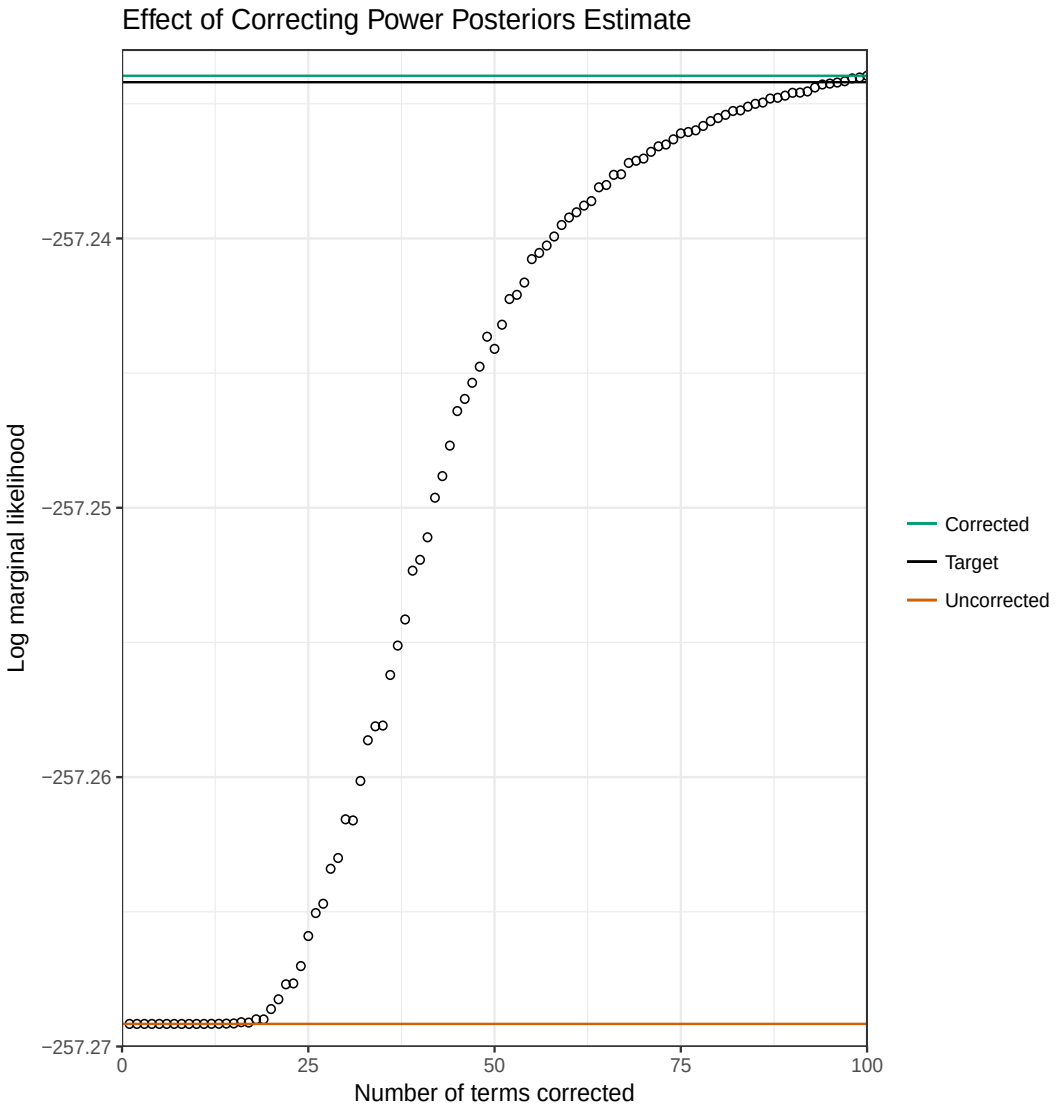



Figure 2-6: Examining the effect of using successively more correction terms in the method of power posteriors. The correction for term $i$ uses an estimate of the variance of the natural logarithm of the deviance at temperature $t_i$ with $t_0 = 1$ and $t_{101} = 0$.

## 2.11 Conclusions

From the first example, in section 2.9, it can be seen that the harmonic mean estimator is a poor estimator that should not be used to obtain a meaningful estimate that could be used in model selection decisions with any certainty. This by now is an unsurprising result as most other reviews of methods in the literature have found this to be the case. Therefore we discard this method as a possibility for model selection with pollution models. Annealed importance sampling will also not be pursued as a method either in the remainder of this thesis, not because it is a terrible estimator, but because it has a noticeably lower precision than the other thermodynamic integration methods for the same cost in computation time. This leaves four methods which will be considered for estimating the marginal likelihood: Laplace's method, Chib's method, steppingstone sampling, and the method of power posteriors.

The second example poses no problem for these methods except for Chib's, which is perhaps a problem just when using Chib's method for estimating the marginal likelihood using a Metropolis-Hastings sampler. If Gibb's sampling is possible then perhaps Chib's method will still perform well. But since in many cases marginal densities for all the parameters may not be readily available this could be a significant problem. In particular for some variogram models, which will be discussed in Chapter 3, the parameters which control the smoothness of the model do not have marginal densities of closed form. Therefore we also leave Chib's method on the wayside here and continue onwards with the three remaining methods.

As both examples demonstrate, when it is feasible to use Laplace's method and the problem is relatively low dimensional, then it provides a fast and accurate solution. Laplace's method is also easy to implement presuming that an accurate estimate of the posterior maximum has been found, which should be the case in most Bayesian analyses. The thermodynamic integration methods give an accurate answer as well but just take longer to do so, however they may be able to provide accurate answers when Laplace's method will fail. In addition the thermodynamic integrations are relatively straightforward to implement as they only require a small modification of a Markov chain Monte Carlo sampler to get results, and steppingstone sampling estimates and power posterior estimates can be obtained from the same Monte Carlo samples. Thus going forward these three methods will be the ones considered in the rest of this thesis.

# Chapter 3

# Spatial Statistics

## 3.1 Introduction

After testing different methods for evidence estimation in the previous chapter, we move forward with a view to applying these methods to models used in the area of spatial statistics, in particular to models of pollution. Firstly the type of data will be described, followed by a characterisation of the statistical models used to capture the nature of the data. Then a particular, but flexible, class of model is studied which is commonly used in the area. Finally the best methods for estimating the evidence seen in the previous chapter, namely the method of power posteriors and steppingstone sampling are applied in a simulated example, and their performance is examined.

## 3.2 Spatial data

To set up the framework for a spatial statistical model we first specify what is meant by spatial data, and why new methods are required for inference with such data. Almost all data has location associated with it, but spatial data refers specifically to data that is believed to have an underlying dependence structure associated with distances between locations. Therefore approaches to making inferences from spatial data need to take this dependence into account, since this violates the main assumption needed for classical statistical inference, that there is independence between observations.

Denote a collection of observations as a vector $\boldsymbol{s}$, contained in a domain $D$ which is a subset of $\mathbb{R}^d$ and can be fixed or random. In many applications the domain of the spatial data is a subset of $\mathbb{R}^2$. We define a function $Z$ on the domain so that $Z(\boldsymbol{s})$ is the set of observations at the set of locations $\boldsymbol{s} \in D$. The spatial process is then denoted

as

$$\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\} \tag{3.2.1}$$

The notation and classification that we are using here is as found in [Cressie, 1991] and was also used in [Schabenberger and Gotway, 2004a]. Also as was done in those books and others in the area, [Shaddick and Zidek, 2015] for example, we categorise spatial data into three main types which are: geostatistical data, lattice data, and point pattern data. These are principally differentiated between by the properties of $D$, the domain.

For geostatistical data $D$ is a fixed $d$ dimensional subset of $\mathbb{R}^d$ of positive volume, so the domain is continuous and non-stochastic, we imagine that there is continuous variation over an area and between any two observations we can take another observation. The primary example of geostatistical data is distribution of ore underground, since this is where geostatistical techniques were first developed. There are many other examples in diverse fields though, such as meteorology, hydrology, oceanography and agriculture. Importantly, pollution data is geostatistical data since pollution is modelled as varying continuously over an area, which means that we will be primarily concerned with statistical models for geostatistical data.

Lattice data, occurs when the domain $D$ is fixed and consists of a collection of countably many points in $\mathbb{R}^d$. Each point observation is always representative of a region, often the result of spatially aggregating over the region, and the location of the point is usually the centre of the region. It may be though that the location represents where the data was collected, such as a school, rather than the centre of the area, which is important to consider as it may affect an analysis. This type of data leads to more choice of distance metric as instead of the Euclidean metric, for example a metric based on examining connectivity may be more useful, so that the distance between two points is defined in terms of the minimum number of borders that would have to be crossed to get from one point to the other. Lattice data occurs when there are data for regions such as counties, and each observation could be an average or count of events over a county, common examples are health data such as number of mortalities, incidences of a disease or economic data such as house prices or average GDP. For example see [Sain and Cressie, 2007], which contains data related to environmental equity in South Louisiana. In [Aukema et al., 2008] we see lattice data related to outbreaks of mountain pine beetle in Western Canada, a subset of this data is also used in [Zhu et al., 2010] where selection of models for lattice data is considered.

Point pattern data is where $D$ is a point process in $\mathbb{R}^d$, and the domain changes

from one observation of the process to the next. This means the domain is stochastic, unlike for the other two categories where the domain is predetermined. Studies of locations of lightning strikes is an example where the data is treated as a point process and there are many examples in the literature of this, see [Podur et al., 2003], [Van Wagtendonk and Cayan, 2008], or [Wang and Anderson, 2011]. Earthquake epicentres are another example of point process data, in particular there is significant spatial correlation in the location of aftershocks following a large event [Jagla, 2010].

## 3.3 Spatial autocorrelation

The key characteristic of a spatial model is the spatial autocorrelation function for the model, which reflects the autocorrelation of the spatial data. Spatial autocorrelation refers to correlation between values of the same attribute at different geographical locations at the same point in time. Temporal correlation may be considered, but here we consider a set of spatial data at only one instance in time. The autocorrelation function should reflect the fact that observation values at two locations are stochastically dependent. In particular it is expected that near points will be more closely related than points which are separated by a greater distance, an intuitive notion neatly encapsulated in Tobler's first law of geography "everything is related to everything else, but near things are more related than distant things." [Tobler, 1970]. We will next consider two ways to capture this law, through the *covariance function* or *semi-variogram*.

### 3.3.1 Covariance function

We define a continuous covariance function acting on two elements, $s_1$ and $s_2$, of the spatial domain $D$, by

$$C(s_1, s_2) = \mathrm{Cov}[Z(s_1), Z(s_2)] = \mathrm{E}[\{Z(s_1) - \mu(s_1)\}\{Z(s_2) - \mu(s_2)\}]. \tag{3.3.1}$$

It is possible and often valuable to make further assumptions about the behaviour of the covariance function, known as stationarity assumptions.

### 3.3.2 Stationarity and isotropy

Using the definition from [Schabenberger and Gotway, 2004b], a process is strongly, or strictly, stationary if the spatial distribution is invariant under translation of the coor-

dinates, so that for all $k \in \mathbb{N}$, $z_i \in \mathbb{R}, h, s \in D$,

$$\mathbb{P}(Z(s_1) < z_1, Z(s_2) < z_2, ..., Z(s_k) < z_k) = \\ \mathbb{P}(Z(s_1 + h) < z_1, Z(s_2 + h) < z_2, ..., Z(s_k + h) < z_k). \tag{3.3.2}$$

This is quite a restrictive assumption so in most cases we instead make the weaker assumption that the spatial process is *second-order, or weakly, stationary*, which implies that the spatial process has a constant mean $\mu$ throughout the domain, $\mathrm{E}[Z] \equiv \mu$, and that the covariance function is a function of the distance between points only, and not of the location itself, so for two points $s_i$ and $s_j$, with a distance $d_{ij}$ between them we have that

$$\mathrm{Cov}(s_i, s_j) \equiv C(d_{ij}). \tag{3.3.3}$$

This means that the covariance between two values depends only on the distance, regardless of where the two values are located. As another consequence of second-order stationarity $\mathrm{Var}[Z(s)]$ is constant everywhere in the space. Note that strong stationarity implies second-order stationarity as long the covariances are well defined everywhere in the domain, but that second-order stationarity does not imply strong stationarity. An example of a process that is strongly stationary but not second-order stationary would be a process where at each point it obeys an identical Cauchy distribution. For a process that is second-order stationary but not strongly stationary, consider a one dimensional process defined over the natural numbers, with one distribution for the odd numbers and another distribution for the even numbers. Assume furthermore that all points are independent so the covariance is zero at any point and that the means and variances of the distributions are the same then it is a weakly stationary process but it is clearly not strongly stationary.

A process is non-stationary if it is neither strictly stationary nor weakly stationary. We do not discuss all properties of stationarity here, for a more detailed listing, see for example [Chatfield, 2016] or other books in the time series literature.

Another important feature of spatial processes is isotropy, which in simple terms is whether direction matters for the spatial process or simply distance. If the covariance is the same regardless of the direction of $d_{ij}$ then the process is called isotropic. Explicitly, a second-order stationary process is *isotropic* if

$$\mathrm{Cov}(s_i, s_j) \equiv C(|d_{ij}|). \tag{3.3.4}$$

A process that is not is not isotropic is *anisotropic*. An example where anisotropic models are relevant is in oceanography, eg. [Worm et al., 2005]. Anisotropy is important to consider in pollution models. If there are strong prevailing winds, there may be a stronger correlation between points in the direction of the wind than between points the same distance apart but not in the direction of the wind.

Whether we assume second-order stationarity is dependent on whether we believe that there is an underlying global mean and variance. Often it is not clear whether second-stationarity holds so we usually begin with the simpler model, assuming that it does hold, then increasing the complexity of the model if it fails to describe the data adequately. Similarly isotropy will be assumed to begin with, unless it can be easily pertained from the data that there is higher covariance in a particular direction. See for an example Figure 3-1, clearly if we are presented with data such as that in the left graph then the anisotropic nature needs to be taken into account. Unfortunately we also have to consider that in practical applications that by increasing the complexity of the model it can quickly become difficult to adequately fit a model, so it is not always possible to include everything that we would like in the model. Although of course on the other side of the coin we do not want to define a model that can be fit but is useless.

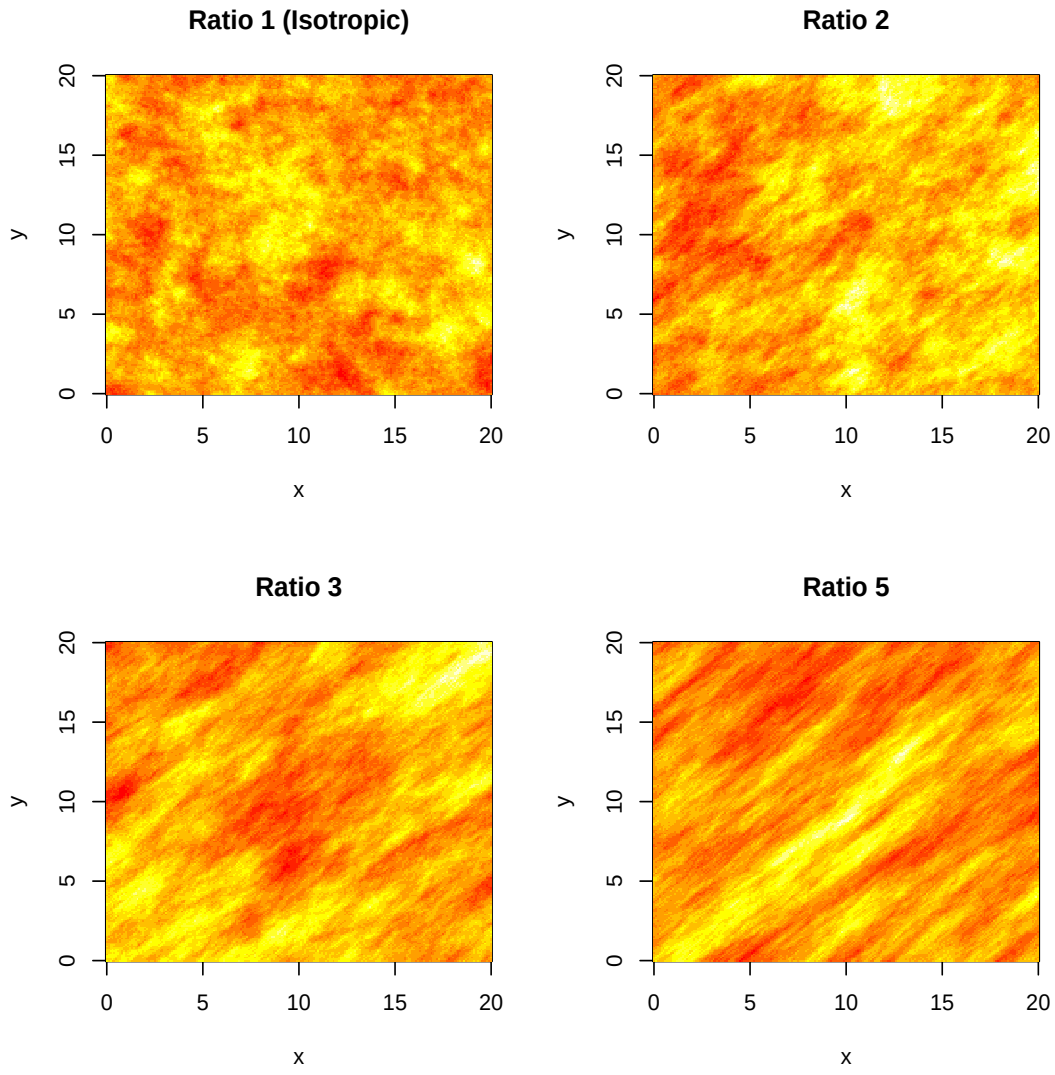



Figure 3-1: Simulations of a spatial field, the top left being an isotropic process, the other three are anisotropic processes with the covariance being larger at an angle of $\pi/4$ radians from the y axis. The anisotropy here is defined by the ratio of the covariance along the specified direction, to the covariance along the y axis.

### 3.3.3 Semi-variogram

For two arbitrary vectors in the domain, $s_1$ and $s_2$, the *variogram* is defined as

$$\begin{aligned}\nu(s_1, s_2) =& Var[Z(s_1) - Z(s_2)] \\ =& Var[Z(s_1)] + Var[Z(s_2)] - 2Cov[Z(s_1), Z(s_)1)]\end{aligned} \tag{3.3.5}$$

If the process is weakly stationary, then this simplifies to

$$\nu(s_1, s_2) = 2\sigma^2 - 2C(h) = 2(C(0) - C(h)), \tag{3.3.6}$$

where $h = |s_2 - s_1|$, the distance between $s_1$ and $s_2$. Thus for a weakly stationary process the variogram will be considered as a function of $h$, $\nu(h)$. The *semi-variogram* is half the variogram, which is defined here for a weakly stationary process as

$$\gamma(h) = C(0) - C(h) \tag{3.3.7}$$

If it exists, the limit of the semi-variogram, $lim_{h\to\infty}\gamma(h)$, is called the *sill*, and is the maximum variance, as when $h \to \infty$ the covariance $C(h) \to 0$. For second-order stationary models so the sill is equal to $C(0) = Var(\mathbb{Z})$. Since the sill may only be reached asymptotically it may often be useful to consider the distance at which the semi-variogram attains a certain proportion of the sill, commonly $95\%$. Referred to as the *range* of the semi-variogram, it is a simple, but practical, approximation to the distance when the semi-variogram has effectively stopped increasing, or equivalently when the covariance between two points has almost stopped decreasing.

An additional feature may be added to a model defined by a covariance function or semi-variogram, which is termed the *nugget*. It is a variance term for the difference of observations at same sites, the variance at zero distance. It is defined to be $lim_{h\downarrow 0}\gamma(h)$, recall $\gamma(0) = 0$ by definition, so the nugget is a discontinuity in the semi-variogram at the origin. This term can be seen commonly used in the geostatistical literature, for example see [López-Granados et al., 2002] or
[Cambardella and Karlen, 1999].

## 3.4 Gaussian Random Fields

### 3.4.1 Random fields

Here we quote the definition of a *random field* from [Abrahamsen, 1997], where many useful definitions and results related to, in particular, Gaussian random fields are compiled.

"Given a probability space $(\Omega, \mathbb{F}, \mathcal{P})$ and a parameter set, $T$, a random field is a real valued function $Z(\mathbf{t}, \omega)$ which, for every fixed $\mathbf{t} \in T$ is a measurable function of $\omega \in \Omega$."

This then is the base structure of a spatial process, note that onwards the existence of the underlying probability space will be implied and not explicitly stated again. A spatial process is therefore a special case of a random field, where we take the parameter set of the random field to be $\mathbb{R}^2$, most commonly, although it may be $\mathbb{R}$ or $\mathbb{R}^3$. The other defining characteristic for a spatial process is a spatially dependent covariance structure on the random field and where the parameter set consists of geographical, or coordinate, points.

### 3.4.2 Gaussian random fields

Gaussian random fields are one of the most important types of random field to discuss as they are widely used in applications in the literature. A random field is a *Gaussian random field* if for all $k \in \mathbb{N}$, for $i = 1, ..., k,\ t_i \in T, Z(t_i) \in \mathbb{R}$, $(Z(t_1), Z(t_2), ..., Z(t_k))$ is equivalent to a $k$-dimensional multivariate Gaussian distribution.

We return now to using $S$ to represent the parameter set rather than $T$ to signify that we are considering parameter sets that have a spatial context. If realisations at $n$ distinct locations $s_1, s_2, ..., s_n \in S$ are denoted as

$$\mathbf{y} = (y_1, y_2, ..., y_n) = (Z(s_1), Z(s_2), ..., Z(s_n)), \tag{3.4.1}$$

then for a Gaussian random field the likelihood, $f(\mathbf{y}|\boldsymbol{\theta})$, is equal to

$$\frac{1}{\sqrt{(2\pi)^n |\boldsymbol{V}|}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\intercal \boldsymbol{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right), \tag{3.4.2}$$

$\boldsymbol{\mu} \in \mathbb{R}^n$. For this to exist, $\boldsymbol{V}$ must be a valid covariance matrix, which is equivalent to the statement that $\boldsymbol{V}$ must be positive semi-definite. Therefore the covariance function $C$ must be a positive semi-definite function. In fact for Gaussian random fields, positive

semi-definiteness of the covariance function is a sufficient and necessary condition for the likelihood to be well defined [Abrahamsen, 1997].

An aside is made here about *Gaussian Markov random fields* (GMRF), since they are currently an important topic in the area, and their relation to Gaussian random fields. A GMRF, loosely speaking, is a Gaussian random field on a discrete graph where the full conditional distribution at a node is equal to the distribution conditioned on just the neighbouring points. Equivalently, in a GMRF two distinct nodes are conditionally independent given all other nodes if there is no edge directly connecting the two nodes, known as the pairwise Markov property [Rue and Held, 2005]. This construction leads to a sparse precision matrix, permitting the use of fast numerical methods which exploit the sparsity. Because of this there has been work made in this area to use GMRFs to approximate continuously indexed Gaussian random fields with applications to spatial statistics amongst other areas [Rue et al., 2009]. Then in [Lindgren et al., 2010] an explicit link is shown between certain Gaussian fields and GMRFs by using an approximate solution to a corresponding stochastic partial differential equation.

## 3.5 Covariance functions

### 3.5.1 Matérn class

At the end of the previous section it was stated that our choice of $C$, the covariance function, must be positive semi-definite. This leads to some methods to construct valid covariance matrices to use within a spatial model. A covariance function $C$ is a positive semi-definite function if, for any set of real numbers $a_1, a_2, ..., a_n$ and any locations $s_i$ and $s_j \in S$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(s_i - s_j) \geq 0. \tag{3.5.1}$$

This implies that $C$ has a certain spectral representation through which Matérn (1986) developed a flexible family of covariance functions to satisfy this condition, the Matérn class. This class is always second-order stationary, and isotropic if the distance metric is the Euclidean metric, which will be assumed throughout. One parametrisation of this class of functions is

$$C(d) = \sigma^2 \frac{1}{\Gamma(\nu)} \left(\frac{\phi d}{2}\right)^{\nu} 2K_\nu(\phi d) \qquad \nu > 0, \phi > 0, \tag{3.5.2}$$

where $K_\nu$ is the modified Bessel function of the second kind of order $\nu$, where $\nu > 0$. The parameter $\phi$ controls the range of the spatial dependence and $\nu$ controls the smoothness of the function, specifically $C(0)$ is $\nu$ times differentiable at $h = 0$. See figure 3-2 for plots of varying values of $\nu$.

If $\nu = m+1$, where $m \in \mathbb{N}$, then (3.5.2) reduces to a polynomial of order $m$ in $d$ and $\exp(-d)$. In particular at $\nu = 1/2$, the Matérn function is equivalent to the exponential covariance function,

$$C(d) = \sigma^2 \exp(-|\phi d|). \tag{3.5.3}$$

This is a very widely used covariance function throughout the spatial statistics literature.

The other ubiquitous covariance function obtained from the Matérn covariance function is the *Gaussian covariance function*. Taking the limit $\nu \to \infty$ in (3.5.2), it becomes

$$C(d) = \sigma^2 \exp(-(\phi d)^2) \tag{3.5.4}$$

The name can be misleading as a random field with a gaussian covariance function does not necessarily follow a Gaussian distribution. It is the smoothest form of the Matérn function, as can be seen from the simulated data in the bottom right graph in figure 3-3. It is perhaps often too smooth to accurately capture real data which is commonly quite jagged.

### 3.5.2 Powered exponential covariance function

There is also another range of functions which encompasses the exponential and gaussian covariance function, and that is the powered exponential covariance function, which was utilised in the design of deterministic computer experiments
[Sacks et al., 1989]. It is defined as

$$C(d) = \sigma^2 \exp(-|\phi d|^k) \quad \text{where} \quad 0 < k \leq 2, \tag{3.5.5}$$

which coincides with the regular exponential model for $k = 1$ and the gaussian model for $k = 2$. Similarly to $\nu$ in the Matérn class, $k$ controls the level of smoothing in the model, see figure 3-3 for examples of data generated using different values of $k$.

### 3.5.3 Spherical covariance function

The *spherical covariance function* is defined as

$$C(d) = \begin{cases} \sigma^2 & d = 0 \\ \sigma^2(1 - \frac{3}{2}(|d|\phi) + \frac{1}{2}(|d|\phi)^3 & |d| \leq \phi^{-1} \\ 0 & \text{otherwise} \end{cases}$$

From this definition of the function notice that there is a cutoff at $d > \phi^{-1}$, which is implying for points that are a distance further apart than a certain value, the covariance between them is negligible. This intuitively makes sense for certain systems, and also means that the covariance matrix will generally be sparse. This can be a practical advantage as computation time is significantly reduced if sparse matrix algorithms are utilised effectively. The spherical covariance function has been used in the environmental and geological sciences amongst others.

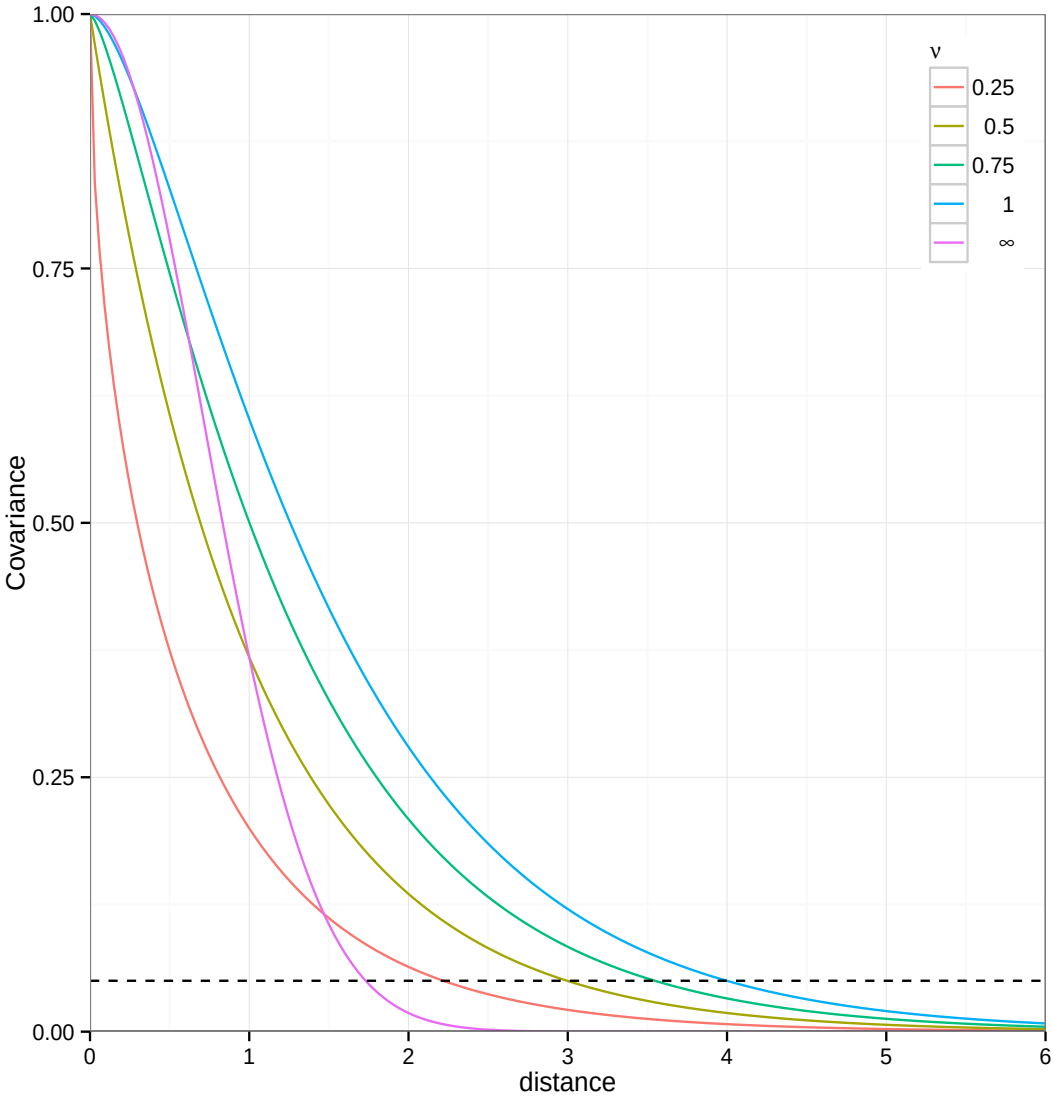



Figure 3-2: Plots of the Matérn function with differing values of $\nu$ which affects the behaviour of the function.

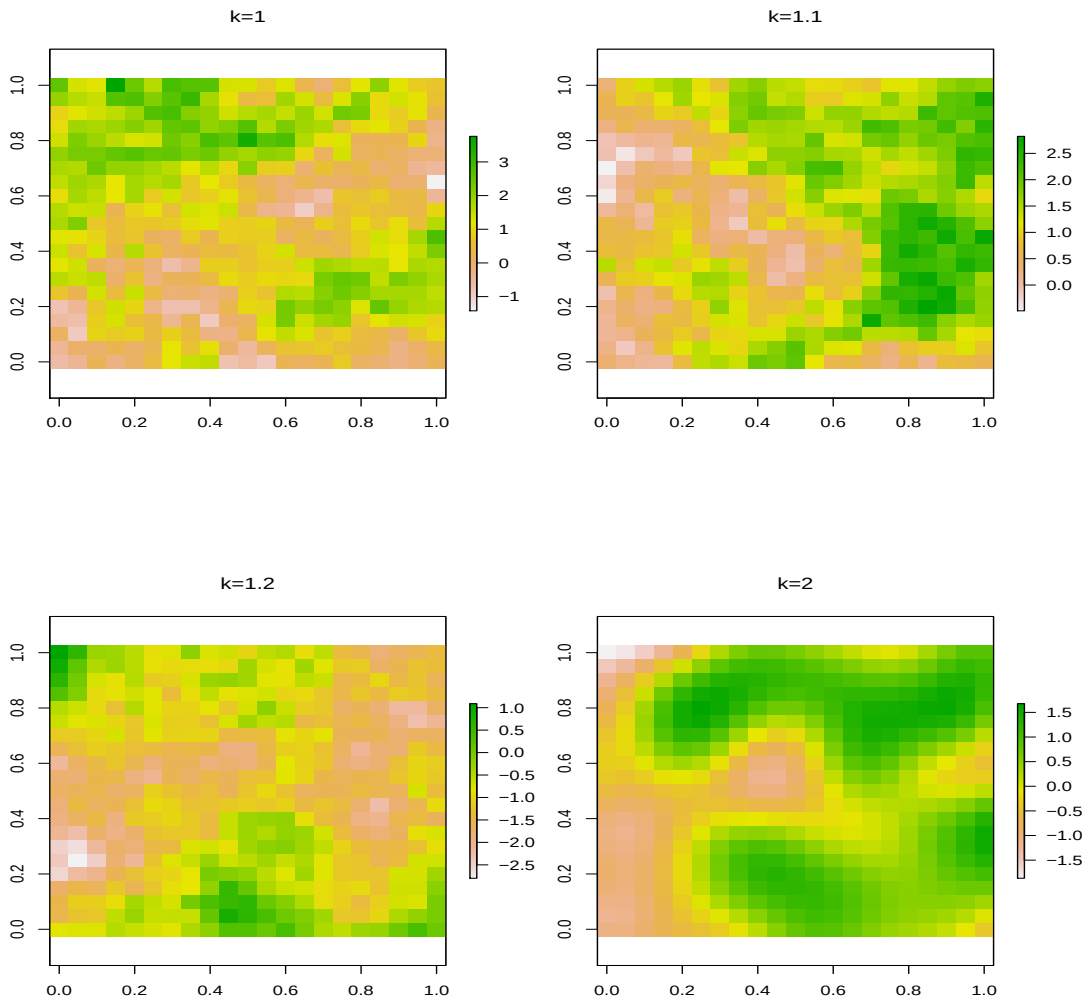



Figure 3-3: Plots of data generated using a powered exponential covariance function for different values of the smoothness parameter, $k$, with the other parameters fixed. Smoothness increases as $k$ increases, from $k = 1$ the exponential covariance function, up to $k = 2$, the gaussian covariance function.

## 3.6 Numerical computation of multivariate normal densities

Before we consider an example of a GRF it will be useful to consider how to efficiently calculate the density function (3.4.2). The bottleneck in the MCMC methods is the evaluation of the likelihood which has to be carried out at each new set of proposed values of the parameters, which requires significant computational effort when the size of the matrix increases, more so depending on the technique used. Straightforward evaluation of (3.4.2) will simply not work in some cases, for example when the determinant of the matrix is too small and the computation of the determinant underflows.

An arithmetic operation is said to overflow, or underflow, when the result is greater than the maximum value, or lower than the minimum value, that can be represented by a given number of bits. In the case of calculating the determinant of a covariance matrix $M$, in double precision, $|M|$ will be evaluated as zero even when $M$ is of reasonably low dimensions, perhaps for as low as thirty depending on the exact choice of $M$. This problem can often be avoided by working on the logarithmic scale, but in this example it is not a direct solution, the logarithm and determinant operations do not commute. It is necessary for instance to first decompose the matrix into a form where the determinant consists of a product, then the logarithm of the determinant can be calculated as a sum of logarithms. See (3.6.2) for an explicit example of this.

Another solution to the above problem would be to increase the machine precision, instead of working with $32$ bit numbers, use $64$ bit numbers. However this is a poor solution as it is simply pushing back the problem, if the dimensions of the matrix increase then the problem resurfaces again. Not only this but the solution through decomposition is much faster, as will be shown in section 3.6.1.

The other problem of calculating and storing the inverse is both time and memory consuming. Although the latter only becomes a problem when the matrix is extremely large, say several thousand rows and columns, with spatial-temporal data this can often occur. On top of that having to multiply the inverse is also a costly operation for a large matrix. As what we want to evaluate is

$$(y-\mu)^{\intercal}\Sigma^{-1}(y-\mu),$$

it is not necessary to form the inverse at any point. Instead, by first factorising the matrix and then exploiting matrix properties, computation costs can be drastically reduced.

The conclusion here is that even if it is possible to obtain the determinant and the inverse directly, this is often not the best solution, it is important to consider carefully the approach to the problem. In the next two subsections we will look at a matrix decomposition method, relevant to covariance matrices, and proceed to demonstrate how this can be utilised to expedite the calculation of multivariate normal densities.

### 3.6.1 Cholesky Factorisation

For a positive semi-definite matrix, there exists a Cholesky decomposition, defined as $R^\intercal R$ where $R$ is an upper triangular matrix called the Cholesky factor of the matrix [Watkins, 2004]. As stated before in section 3.5.1, covariance matrices by definition are positive semi-definite, and in particular with a covariance matrix defined by the Matérn function we have that the covariance matrix is positive definite as long as the points are distinct. So we can write such a covariance matrix $\Sigma$ in the aforementioned manner, which will be useful for calculating the multivariate normal density. Cholesky decomposition is definitively faster as a general method for decomposing covariance matrices requiring $O(n^3/3)$ flops, see for example [Watkins, 2004].

Cholesky decomposition may be used to greatly increase the speed of computing the likelihood over naively directly calculating $\exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$ using the inverse. As the Cholesky decomposition exists, $\Sigma$ can be decomposed into $R^\intercal R$, where $R$ is the upper triangular Cholesky factor. Then the term we are interested in, $(\mathbf{y}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$, becomes $a^\intercal a$ where $a = R^{-1}(y-\mu)$. Next we need to solve for $a$ in

$$Ra = y - \mu$$

As the Cholesky factor is upper triangular, we do not need to evaluate $R^{-1}$ and multiply by $y-\mu$ , we can use backward substitution to solve for $a$. In addition, from this method we get the determinant extremely cheaply, as $|\Sigma| = |R^\intercal||R|$ so that

$$|R^\intercal| = |R| = \prod_{i=1}^{n} R_{ii} \tag{3.6.1}$$

and therefore on the logarithmic scale,

$$\log|\Sigma| = \log(|R|^2) = \sum_{i=1}^{n} 2\log(R_{ii}). \tag{3.6.2}$$

### 3.6.2 Implementation for an exponential model

In this section the implementation of the approach discussed in the preceding section is shown, in general and also considering a range of discrete values of $\phi$ which is useful for example when the prior on $\phi$ is a discrete uniform distribution. Before that though choice of priors will be considered, and why a discrete uniform prior on $\phi$ in the model may be beneficial in practice.

In a Bayesian setting one is confronted with the difficult choice of prior selection on parameters, which allows potentially subtle errors to be made. Fortunately by definition, some parameters have a significance in the model which informs the choice of prior. For an exponential covariance, $\phi^{-1}$ has a relation to the range, see section 3.3.3. This can be used to bound the prior on $\phi$ since we expect the range to fall within certain distances depending on the data, there is no point considering values of $\phi$ which give an effective range greater than the maximum distance between any two points for example. Similarly it is no use considering values of $\phi$ where the effective range is smaller than the minimum distance between any two points. Therefore it is often the case that bounded priors are chosen for $\phi$, see [Finley et al., 2007] or [Diggle et al., 1998] for examples. As noted in [Shaddick and Zidek, 2015] however, this approach can be very computationally demanding and so here we consider the possibility of choosing a discrete uniform distribution as the prior, thereby restricting possible draws to a finite number which will reduce the amount of calculations that need to be made for MCMC sampling.

Given a finite number of values for $\phi$ we only have to calculate a much smaller number of Cholesky factors overall, one for every value of $\phi$, before running the chain, rather than having to do this at every iteration. This is because, if $\sigma^2 V = \mathbf{\Sigma}$ then $|\mathbf{\Sigma}| = \sigma^{2k}|V|$ and $\mathbf{\Sigma}^{-1} = \sigma^{-2}V^{-1}$. So we store the Cholesky factors of all the covariance matrices possible from the discrete range of $\phi$, and scale them by $\sigma$ when necessary, thus reducing the total computation time of each estimate overall. As $\mu$ and $\sigma$ are varying continuously at each iteration of the MCMC algorithm, the back solving still must be done within the MCMC algorithm, but the factorisation does not have to be done at every step so it is still a significant reduction in computation time.

Now we examine how the likelihood in the exponential model can be calculated explicitly. Note that we If we have $\phi_s,\ s = 1, 2, .., m$, and a distance matrix $\mathbf{D}$, take just one $\phi_s$, and consider the calculation of $\mathbf{\Sigma}^{(s)}$, where $\mathbf{\Sigma}^{(s)}_{ij} = \sigma^2 exp(-\phi_s D_{ij})$, then in R the following code will calculate the determinants and Cholesky factor of $\mathbf{\Sigma}^{(s)}/\sigma^2$

```
Sigma = exp(-D*phi)
```

```
R = chol(Sigma)
detS = sum(log(diag(R)))
```

At each step of the MCMC, if mu is the current value of $\mu$ and sigma is the current value of $\sigma$, and $k$ is the dimension of $\mathbf{x}$, then the log likelihood is given by:

```
B = backsolve(r = R, t(x)-mu,transpose = T)
ll = -(k/2)*log(2*pi) - detS  -k*log(sigma)
         - 0.5*(t(B) %*% B)*(sigma^(-2)).
```

From this code you can see that to factor $\sigma$ back into the likelihood involves adding $k\log(\sigma)$ to $|\Sigma|$ and subtracting sigma in the last term, the one that involves the inverse of $\mathbf{\Sigma}$.

### 3.6.3 MCMC sampling

Given that samples are required from the power posteriors, the question is which MCMC algorithm to use to generate such samples. Metropolis-Hastings was chosen as it is the most general method of sampling from posterior distributions and hence also works for sampling from the powered posterior distributions. The standard random walk Metropolis-Hastings algorithm can be very inefficient in high dimensions but since the number of parameters is relatively low for the models here this problem should not be seen. The alternative would be to use a Hamiltonian Monte Carlo algorithm which would explore the state space more efficiently with proper implementation, which would mean a significant reduction in the time taken to form an estimate. For an introduction to Hamiltonian Monte Carlo see [Betancourt, 2017]. A Metropolis-Hastings was used however due to the simplicity of implementation which keeps the focus on the method of estimating the marginal likelihood rather than the sampling algorithm.

In particular for the examples in this work in Chapter 3 and Chapter 4, a random walk Metropolis-Hastings algorithm is used to sample the state spaces. The procedure used is summarised at each step by the following:

1. Propose a change in parameter from a Gaussian distribution centred at the current position (or a simple symmetric random walk if the parameter is discrete).

2. Accept the move with probability equal to the likelihood evaluated at the new proposed point divided by the likelihood evaluated at the previous point, $\pi(y|\theta^{(new)})/\pi(y|\theta^{(old)})$.

3. Repeat steps 1 and 2 for all parameters.

4. Iterate.

The variance of the proposal jumps are tuned to get desirable acceptance rates, which means that the chain moves frequently but also moves far enough that the chain will explore the sample space well and not linger in one region of the distribution.

## 3.7 Simulated example

We move on to look to see how these methods could be applied in practice with a simulation study in R with data generated from a Gaussian random field with an exponential covariance function. This model is assuming that the data follow a multivariate normal distribution with the covariance given by (3.5.3). Explicitly, this means that the elements of the covariance matrix, $\mathbf{\Sigma}$, are defined for $i, j = 1, 2, ..., n$ as

$$\mathbf{\Sigma}_{ij} = \sigma^2 \exp(-\phi d_{ij}). \tag{3.7.1}$$

Data was randomly generated using a uniform twenty by twenty grid on a unit square giving four hundred and forty one points of data in total. The parameters were: $\mu = 0$, $\phi = 0.5$, and $\sigma^2 = 1$. The data is shown in Figure 3-3 as an example of the powered exponential covariance model with $k = 1$.

The models that will be compared to (3.7.1) will incorporate a powered exponential covariance function, defined as $\mathbf{\Sigma}_{ij} = \sigma^2 \exp(-(\phi d_{ij})^k)$, where $0 < k \leq 2$. We will compare models with different fixed values of $k$. The values of $k$ chosen were: $k = 1$, $k = 1.1$, $k = 1.25$, $k = 1.5$, and $k = 2$. Recall that for $k = 1$ we have the standard exponential covariance function, and for $k = 2$ we have the gaussian covariance function.

Weak normal priors were chosen for the mean $\mu$ and also for $\log \sigma$, centred on the generating values. The parameter $\phi$ was given a discrete uniform prior with bounds chosen which represent low or very high spatial correlation. Trace plots of $\phi$ did not indicate that these bounds were restrictive, that is the chains did not reach these boundaries in testing. There were four hundred values evenly spaced from $0.1$ to $8$. The density of the values of $\phi$ for the discrete prior in the chosen range was selected to give desirable acceptance rates for proposed moves during the Metropolis-Hastings runs.

The estimates of the evidence were made with the stepping stone method and the power posteriors method, detailed in sections 2.7 and 2.8, with ten repetitions made

with each. The estimates from the power posterior method and the stepping stone sampling method use the same samples within each repetition, the only difference being that the stepping stone method does not use samples from the temperature $t = 0$.

To decide how many temperatures to use and how many samples to take at each temperature, preliminary runs were done with varying numbers of temperatures and samples at each temperature. The sample variance of twenty repetitions for the four different specifications of temperatures and samples used were: 3.99, 0.62, 0.11, and 0.23 respectively, going from left to right in Figure 3-4. Following the same order, computation times in hours were: 12, 24, 48, and 48.

From these results we see that going from twenty five to fifty temperatures greatly increased the precision of the estimates but then after that it was more beneficial to use a higher number of samples rather than further increasing the number of temperatures used. In the case of power posteriors, this signifies that the possible gain from a decrease in numerical error by doubling the number of temperatures used is outweighed by the possible decrease in Monte Carlo error by doubling the samples taken at each temperature. Increasing both was considered to raise computation time to be too high for each estimate, with small gains to be made in the precision. From this it was decided that forty thousand samples and fifty temperatures would be appropriate to use.

### 3.7.1 Simulation study results and discussion

Before the results are discussed it should be noted that the scale in [Kass and Raftery, 1995] is referred to, which has the following categories based on the value of twice the log Bayes factor:

- 0 to 2 - Not worth more than a mention
- 2 to 6 - Positive
- 6 to 10 - Strong
- Greater than 10 - Very strong

The results, shown in Table 3.1 on the natural log scale, came out as might have been expected with the evidence for the model with $k = 1.1$ estimated to have the highest evidence, the models where $k = 1$ and where $k = 1.25$ were estimated as having slightly lower evidence, and as having very similar evidence to each other. This would

be considered positive evidence in favour of $k = 1.1$ based on the aforementioned scale. On the other hand for $k = 1.5$, given the evidence estimated, we would reject this value of $k$ when comparing to the other three values of $k$ since the difference in marginal likelihood on the log scale is greater than five. Given the data was generated using $k = 1$ we would hope that this more extreme value, of $k = 1.5$, would be rejected. As for the other values it is perhaps that there is insufficient data to differentiate between the closer values of $k$. This is simply an example of implementation though so the results can not be read into too much, many more replications would be required for further analysis.

Looking at the estimates it can be noted that the estimates differ slightly between the two methods but within each method the differences between the models follow the same pattern. The difference between the two estimates is mainly due to the differing bias in the two methods. The stepping stone sampling method is biased on the log scale and there is discretisation bias in the method of power posteriors due to the trapezium rule. Both methods have variance induced from the stochastic nature of their estimates. The power posteriors estimate seems to suffer from higher variance when $k = 1.5$, which seems to be caused by extreme values of the likelihood being obtained at the lower temperatures, perhaps indicating that the estimate from steppingstone sampling is less sensitive to the sampling of extreme values.

Table 3.1: Table showing the difference in mean log evidences, relative to the mean log evidence for $k = 1$ estimated by power posteriors. Estimates for the log evidence of three powered exponential covariance models distinguished by fixing the values of $k$, the exponent, to be 1, 1.1, 1.25, and 1.5.

| Method | k | Difference in mean | Standard deviation |
|---|---|---|---|
| Power posteriors | 1 | 0 (baseline) | 0.36 |
| | 1.1 | 1.19 | 0.28 |
| | 1.25 | 0.043 | 0.29 |
| | 1.5 | -5.89 | 1.43 |
| Stepping stone Sampling | 1 | 1.13 | 0.30 |
| | 1.1 | 2.24 | 0.22 |
| | 1.25 | 1.04 | 0.24 |
| | 1.5 | -5.67 | 0.38 |

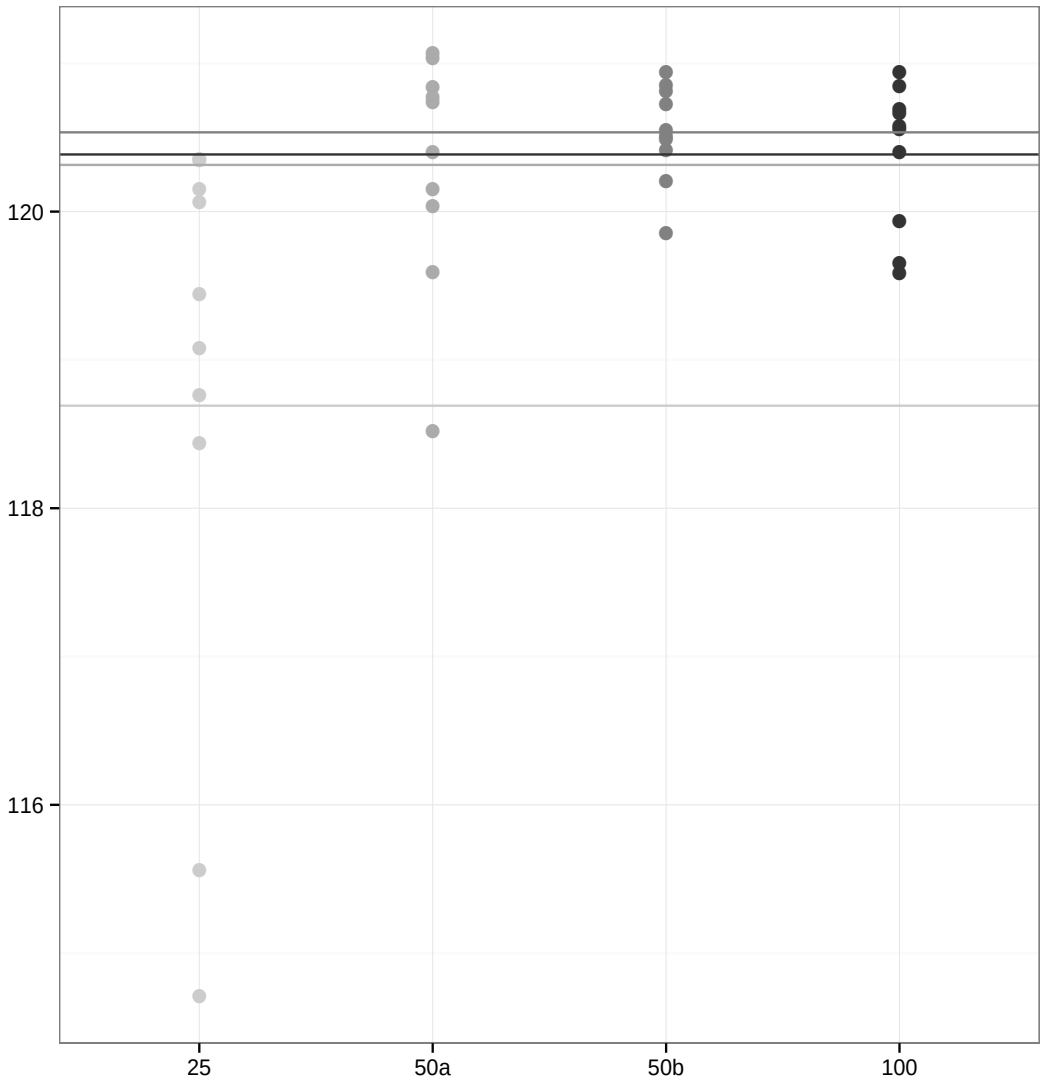



Figure 3-4: Comparison of estimates of the log evidence using the method of power posteriors with varying temperatures, all with twenty thousand iterations at each temperature, except for 50b which had forty thousand iterations at each temperature. The labels on the x-axis correspond to the number of temperatures used, and the horizontal lines correspond to the means of each set of samples.

# Chapter 4

# Bayesian model choice in spatial statistics

### 4.0.1 Introduction

In this chapter some of the current methodology from the literature in regards to model choice will be discussed and how perhaps the marginal likelihood could be used to help inform choice in future studies. This is followed by a case study featuring a dataset from the WHO, regarding pollution levels in Germany and China, specifically particulate matter less than two and a half microns in diameter (PM2.5). The marginal likelihood of alternative models are estimated by three different methods and the results are compared.

### 4.0.2 Model selection criterion

Methods other than the use of Bayes factors for model comparison should be mentioned, two closely linked methods being the *Akaike Information Criterion* (AIC) and the *Bayesian Information Criterion* (BIC). They are related as they are both penalized-likelihood criteria, derived from the likelihood with some penalty related to dimension. The AIC, [Akaike, 1973], measures the deviation of a specified model from the hypothetical "true" distribution. It is based on the maximum likelihood estimate with a bias correction given by the number of parameters in the model. If $\pi(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ is the likelihood function evaluated at the maximum likelihood estimate and $d$ is the number of dimensions then the expression for AIC is $-2\pi(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + 2d$. As can be seen from this expression the AIC does not take priors into account as it was developed from a frequentist perspective rather than a Bayesian one. It has been shown how-

ever that it can be derived in a Bayesian framework under certain prior assumptions, [Burnham and Anderson, 2004].

The BIC was outlined by Schwarz in [Schwarz et al., 1978] as an alternative to AIC, it is defined as $-2\pi(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + d\log(n)$, where $n$ is the number of observations. For a comparison of these methods see for example [Kuha, 2004]. BIC has been shown to be an approximation to Bayes factors [Raftery, 1986] and is often used in lieu of them, BIC being much easier to calculate. The *Deviance information Criterion* (DIC) [Spiegelhalter et al., 2002] is another commonly used criterion in the Bayesian setting which was developed as an alternative to the AIC for models where the number of parameters is unclear, such as hierarchical models. There are of course now many other information criterion, for a somewhat recent overview see [Spiegelhalter et al., 2014].

### 4.0.3 Model selection within the spatial statistics literature

Here is a brief look into some current practices, but first some past quotes on the subject. In Cressie's comprehensive book on statistics for spatial data he notes "A fertile area for future research is in developing sensible model-selection criteria when the data are spatial" [Cressie, 1991]. Almost ten years later Gorsich comments: "Apart from possible a priori knowledge about the process and the user's subjectivity, there is no standard methodology for choosing among valid variogram models like the spherical or the exponential ones", [Gorsich and Genton, 2000]. Now almost twenty years later this is still very much the case. There is a noticeable absence of model comparison in the spatial statistics literature. Often models are selected and fit with no attempt to compare with possible alternative models, meaning alternative covariance structure or alternative covariates. Next is a small sample of some materials from the spatial statistics literature, to examine what methods are discussed or used.

We find no mention of model comparison by any of the methods in [Shaddick and Wakefield, 2002]. In [Finkenst adt et al., 2006], page 98, when fitting a non stationary process, both BIC and AIC are used. For *spBayes*, a software package developed for R that implements Bayesian hierarchical spatial models [Finley et al., 2007], the DIC is the model selection tool of choice. In [Shaddick et al., 2018], a study on world air pollution, INLA was used to fit an approximate model, for model comparison: DIC, R squared statistics, and Cross validation were utilised. Next looking at [Shaddick and Zidek, 2014] where a GMRF model is fit using INLA methods, again DIC is seen to be used, alongside RMSE statistics for model fitting. So from this selection it can be seen that there is no general consensus

on model selection criterion in the literature. DIC seems to be often used but marginal likelihoods, and hence Bayes factors, appear to be absent.

## 4.1 Case Study: World pollution data

In this section data concerning levels of harmful pollutants from locations around the globe will be examined. The data is the same as that which was used in the World Health Organisation's report on global pollution levels [WHO et al., 2006]. Here though subsets of the data will be taken, in particular only levels of $PM_{2.5}$, particulate matter less than two and a half microns in diameter, will be considered, which will be modelled over countries rather than on a global scale. $PM_{2.5}$ is of particular interest due to its links with respiratory diseases, see for example [Abbey et al., 1995].

Firstly a variogram model of $PM_{2.5}$ over Germany will be constructed then models with differing covariates will be proposed and compared using the marginal likelihood. Note that this is simply an exploratory analysis and not a statistical test of the models against each other, the emphasis is on the efficacy of the methods for estimating the marginal likelihood of models in this context. However the results may provide grounds for further analysis of particular models. Germany was selected as it has a high number of monitoring sites which are relatively evenly spread throughout the country.

Then, the data for China will be analysed to provide a contrast, being from a different region of the world, having a less uniformly spread monitoring network, and having more varied levels of pollution. Also there are data for two hundred and five locations in Germany and one thousand one hundred and fifty six locations in China, so the China dataset is significantly larger, providing further comparison of the methods for different size datasets. Graphs of the data from the two countries can be seen in Figures 4-1 and 4-2.

There are five variables that will be considered as potential model covariates, here they will be summarised but for a full explanation of all the data sources see [Shaddick et al., 2018]. The first variable to be considered is an estimate of the pollutant level based on satellite readings, which will be referred to as *SAT*. The second variable is a set of numerically simulated values from TM5-FASST(FAst Scenario Screening Tool) [Van Dingenen et al., 2014] which is related to the TM5 CTM Chemistry Transport Model) [Huijnen et al., 2010]. These values will be referred to as *TM5*. Another variable is a figure that combines information on elevation and distance to nearest urban land surface, *ELDU*, which was formulated and shown to be a signifi-

cant predictor of $PM_{2.5}$ in [Van Donkelaar et al., 2016]. Additionally there is information regarding concentrations of various other chemicals, *SNAOC* (Sodium, Nitrogen, Ammonia, Organic Compounds) and compositional concentrations of mineral dust, *DUST*. Both of these variables are derived from the GEOS-Chem chemical transport model, see again [Van Donkelaar et al., 2016] for details.

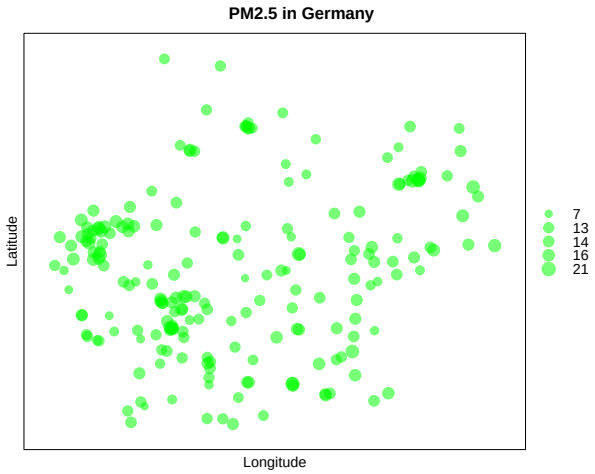



Figure 4-1: Locations and readings of $PM_{2.5}$ in Germany, units are $\mu g/m^3$.

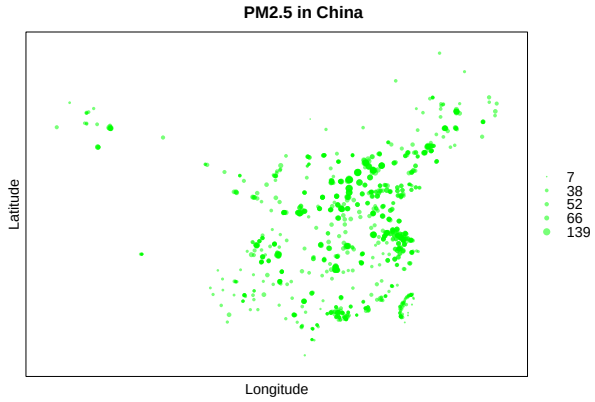



Figure 4-2: Locations and readings of $PM_{2.5}$ in China, units are $\mu g/m^3$.

### 4.1.1 Modelling the data

The spatial model used here to model the level of $PM_{2.5}$, on the logarithmic scale, is an exponential model which as previously discussed is a widely used model in the area. The reason that a more general Matérn class model is not used is that it would allow for too much flexibility in the model, leading to the smoothness parameters being unidentifiable in practice given the amount of data available. Recall from section 3.5.1 that the exponential covariance function is of the form $C(d) = \sigma^2 \exp(-|\phi d|)$. Allowing the $\phi$ parameter to vary was tested but it was found to cause problems with convergence of the Metropolis-Hastings algorithm, causing high autocorrelation in the chain, whether that is due to the amount of data or the structure of the data is unclear. This problem seems to be inherent to the spatial decay parameters in spatial models, which are in general weakly identifiable [Finley et al., 2007]. The other problem with allowing a more flexible model is computation time as it is a limiting factor. When the number of sites is high, around one thousand, each additional variable adds hours to the computation time. Additionally, empirical variograms were generated and it appeared an exponential variogram would be reasonable to attempt to fit to the data, see Figure 4-3. Note that it would be also reasonable to attempt to fit a spherical variogram model to the data judging from the shape of the empirical semivariogram.

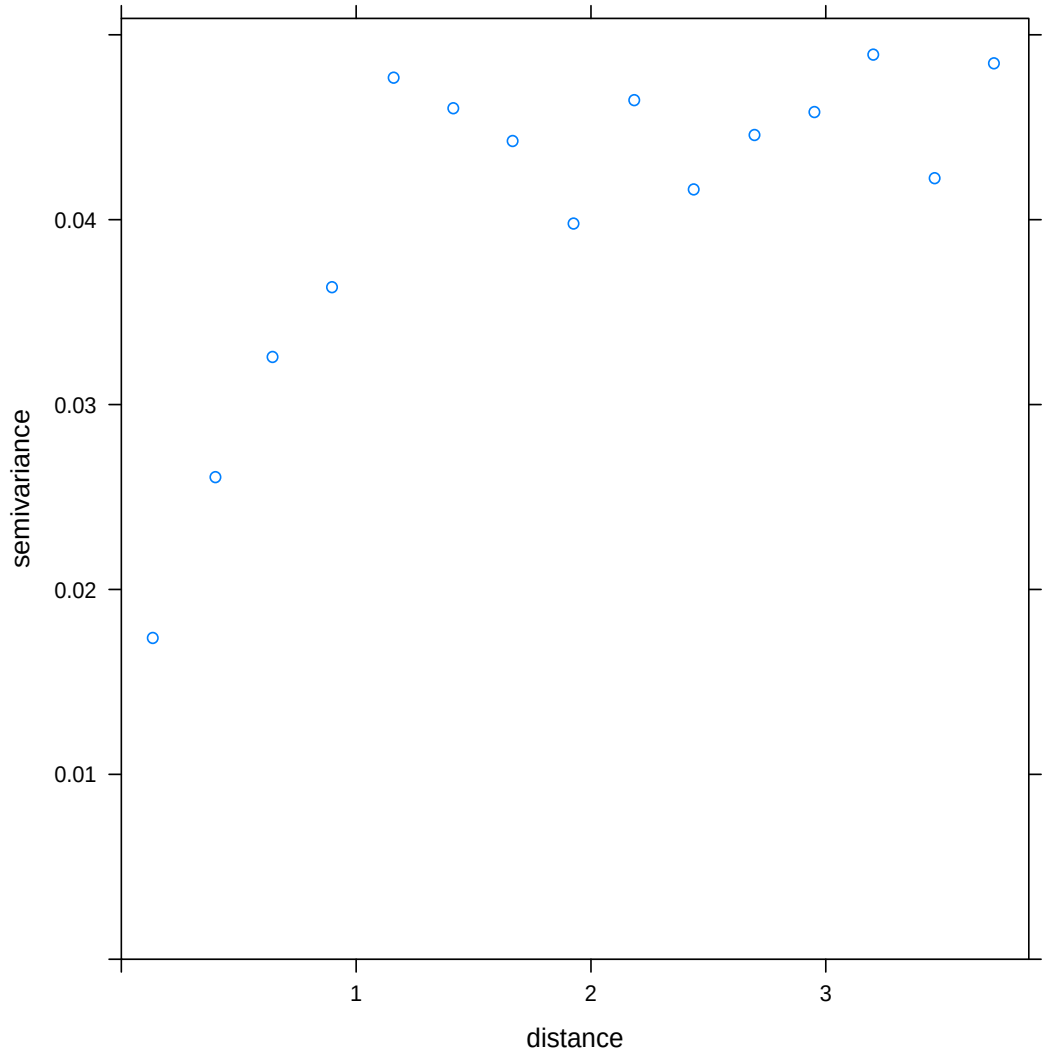



Figure 4-3: An empirical semivariogram of $PM_{2.5}$ in Germany, created using the gstat package [Gräler et al., 2016].

Within the dataset there are 206 sites located in Germany; one site with missing values is excluded for simplification, proper treatment of missing values would require further work beyond the scope of this thesis. This leaves 205 sites with complete data, so the data is of a similar scale to the simulated example in Section 3.7, the two main differences between these examples is that the locations are not uniformly distributed over the area and there are linear covariates introduced in this example. Instead of

there being a constant underlying mean as in the simulated data, we expect the mean level of pollution detected at the ground site to be related to other variables, explicitly we expect the mean of the random field to be $\boldsymbol{\beta_0} + \boldsymbol{X}\boldsymbol{\beta_{1,2,\ldots,k}}$, where each column of $X$ is a predictor variable. The focus of this study will be on different choices of $X$ and calculation of the marginal likelihood for the different models specified by these choices. The models all have the same underlying structure, a Gaussian random field with an exponential covariance function (3.5.3). The likelihood function for model $i$ then is given by

$$\frac{1}{\sqrt{(2\pi)^n|\boldsymbol{C_i}|}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu_i})^\intercal \boldsymbol{C_i}^{-1}(\mathbf{y}-\boldsymbol{\mu_i})\right) \tag{4.1.1}$$

The models are labelled A1, A2, A3, A4 and B1. B1 was only tested for the Germany dataset, due to the low variation in the satellite readings for Germany. The variables in each model are as follows:

- Model A1: POP, SAT
- Model A2: POP, SAT, TM5
- Model A3: POP, SAT, DUST, SNAOC, ELDU
- Model A4: POP, SAT, DUST, SNAOC
- Model B1: POP.

### 4.1.2 Results and Conclusion

First we look at the results for the data relating to Germany. The full results are shown in Table 4.1. Box charts of the results of ten estimations from the method of power posteriors and the results by stepping stone sampling alongside the Laplace method are shown in Figure 4-5. It can be seen that the estimated logarithmic marginal likelihood for models A1, A4, and B1 are very similar, with means within a range of $0.21$ as estimated by the method of power posteriors and even closer from the steppingstone estimates, a range of $0.08$. Thus the marginal likelihood does not distinguish which one of these models is preferable here. Model A2 has the most unfavourable result, being $-4.20$ and $-4.36$ relative to the result for A1 according to power posteriors and stepping-stone methods respectively. Referring to the scale given in [Kass and Raftery, 1995], if we were to calculate the Bayes factor between A1 and A2 based on these results this would be classed as 'strong' evidence in favour of

A1. Specifically it is saying that introducing the TM5 variable to model A2 lessens the power of the model at predicting the data in this instance. Note here that although A1 is nested in A2 the marginal likelihood, the marginal likelihood for A1 is higher than for A2, so the marginal likelihood does not behave like the likelihood function in this sense. The marginal likelihood is also dependent on the chosen prior distributions so prior sensitivity analysis should be enacted before making a model selection based on the Bayes factor.

Model A3 has the highest result compared to the other models. If A3 is compared to any of the other models then this result would indicate strong evidence in favour of A3. It is stressed here however that this experiment was not set up as a rigorous test between these models so the term ”strong evidence” is a little misleading. The results from the Laplace method agree with the results found from the two tempering method except for the result for model A3, it is not as strongly favoured, it would fall into the category of ”positive” evidence in favour of A3 against the simplest models here.

Regarding the variation in the results it can be seen that the sample standard deviation of the estimates of the log marginal likelihood for the simpler models is lower than that of the models which include more variables. For the method of power posteriors, the ten estimates for model A1 have a sample standard deviation of $0.23$ while $A4$ has a relatively much higher sample standard deviation of $0.92$. The same issue is seen with the estimates from the stepping stone sampling method. This indicates that it is perhaps necessary to increase the number of samples at each temperature for models with more parameters to reduce the variance, however it may just be a matter of adjusting the proposals however or a problem inherent to this data. As can be seen here more clearly looking at both results plotted together, Figure 4-6, the variance of the estimates for models A3 and A4 means that there are overlapping values which undermines the result of the mean difference in values. If the purpose of calculating the marginal likelihood is to provide decisive information to decide between two of these models then some effort needs to be made to decrease the variance. From testing in the previous chapter, section 3.7, it was found that increasing the number of temperatures beyond fifty gave little return with regards to increases in precision. Therefore in this case the most effective course to decrease the error should be to take more samples at each temperature. The Laplace method provides a single result and therefore of course does not have an associated error measurement. However the possible error with the Laplace method is in finding the maximum of the posterior. It was found that starting the optimiser at different points would lead to finding local maxima which were clearly not the global maximum. Because of this, using the max-

imal point found during a Metropolis Hastings run as a starting point for a numerical optimiser is recommended, along with a few other starting locations.

Table 4.1: Table showing the log marginal likelihood of different spatial models for $PM_{2.5}$ in Germany, estimated by the method of power posteriors and steppingstone sampling. The models are exponential covariance models which differ by the inclusion or exclusion of covariates, see 4.1.1 for more details.

| Method | Model | Mean | Log Bayes Factor wrt A1 | Standard deviation |
|---|---|---|---|---|
| Power posteriors | A1 | 51.89 | 0 | 0.23 |
| | A2 | 47.69 | -4.20 | 0.57 |
| | A3 | 55.95 | 4.05 | 0.78 |
| | A4 | 52.10 | 0.20 | 0.92 |
| | B1 | 51.89 | -0.01 | 0.12 |
| Stepping stone Sampling | A1 | 51.78 | 0 | 0.24 |
| | A2 | 47.4 | -4.36 | 0.61 |
| | A3 | 54.88 | 3.09 | 0.79 |
| | A4 | 51.76 | -0.02 | 0.93 |
| | B1 | 51.84 | 0.06 | 0.12 |
| Laplace Method | A1 | 51.84 | 0 | |
| | A2 | 47.45 | -4.39 | |
| | A3 | 54.52 | 2.67 | |
| | A4 | 51.33 | -0.51 | |
| | B1 | 51.86 | 0.02 | |

The results for China can be seen in Table 4.2, which greatly differ from the results for Germany. Here A1 and A2 are strongly favoured over the other models. Examining the Bayes factor between A1 and A2 gives ”positive” evidence for A2 in terms of the scale given by Kass and Raftery. It seems here that the information from the TM5 model is valued more than when looking at models for Germany, perhaps due to the increased sparsity of the data. The average distance between sites is lower and there is more uniformity of the spacing of the sites in Germany compared to China.

For models relating to China there is a clear discrepancy in the results from the method of power posteriors with models A3 and A4 which highlights a potential issue in implementation of the method of power posteriors. The exact same procedure and code for calculating these estimates was used throughout this entire example yet only for these two models do such inaccurate values get returned. However upon inspection the problem was seen to be in the correction terms for the power posterior estimate. As was done in Figure 2-6, one estimate was calculated by power posteriors then successively more correction terms were applied to see the resulting effect, see

Table 4.2: Table showing the log marginal likelihood of different spatial models for $PM_{2.5}$ in China, estimated by the method of power posteriors, steppingstone sampling and Laplace. The models are exponential covariance models which differ by the inclusion or exclusion of covariates, see 4.1.1 for more details.

| Method | Model | Mean | Log Bayes factor wrt A1 | Standard deviation |
|---|---|---|---|---|
| Power posteriors | A1 | 289.79 | 0 | 0.23 |
| | A2 | 293.13 | 1.34 | 0.27 |
| | A3 | 276.54 | -13.25 | 1.11 |
| | A4 | 304.92 | 15.13 | 3.39 |
| Stepping stone Sampling | A1 | 289.77 | 0 | 0.20 |
| | A2 | 291.07 | 1.29 | 0.20 |
| | A3 | 271.19 | -18.59 | 0.73 |
| | A4 | 274.65 | -15.12 | 0.38 |
| Laplace method | A1 | 289.80 | 0 | |
| | A2 | 291.07 | 1.27 | |
| | A3 | 271.228 | -18.57 | |
| | A4 | 274.52 | -15.27 | |
| Power posteriors (correction removed) | A1 | 289.01 | 0 | 0.23 |
| | A2 | 290.28 | 1.26 | 0.27 |
| | A3 | 267.75 | -21.26 | 1.11 |
| | A4 | 266.53 | -22.48 | 3.39 |

Figure 4-4. The final terms are far larger than they ought to be, these terms correspond to the power posterior near to $t = 0$, when it is closest to the prior distribution. These issues with the power posterior correction were discussed in Section 2.10.1. One possible solution to this problem would be to have a cut off point in the correction terms, for when they are clearly too large, but there is then the problem of defining and justifying the cut off point. Another possible solution would be to increase the temperatures or change the distribution of the temperatures. As only twenty five temperatures were used here, perhaps with an increased number of temperatures, forty or fifty say, the issue would disappear. This would drastically increase the computation time of each estimate, albeit in a linear fashion. Overall though given that steppingstone sampling seems to suffer none of these drawbacks it seems the better method here. The values of the power posterior estimates were recalculated without the correction and added to Table 4.2, the result for A4 now agrees with the other two methods however A3 and A4 seem to be underestimated compared to the results for the other two methods, possibly due to an increase in the discretisation error.

As seen with the Germany data, for the tempering methods the variance of the estimates seems to increase with the model complexity, although this is not certain as the sample size is so small. It remains to be seen if for a model with significantly more parameters would require similarly large increases in number of samples at each temperature rung. If this is the case it is a significant detractor from these methods as computation time would be further increased. For the Laplace method the danger is in not finding the posterior mode and instead finding a local maximum and thinking that the posterior mode has been found. Starting the search for the maximum at the maximum found during the MCMC sampling appears to fully resolve this issue in this instance.

It has been seen earlier that the method of power posteriors produces results as accurate as steppingstone sampling, but it can be seen here how the method can fail, steppingstone sampling is simply a more robust estimator in the final example. However the Laplace method is clearly the best of the three methods in this example due to matching the steppingstone sampling estimates and taking far less time to obtain those results. As perhaps should have been expected, the form of the likelihood and the relatively low number of parameters makes for an ideal situation for approximation by the Laplace method.

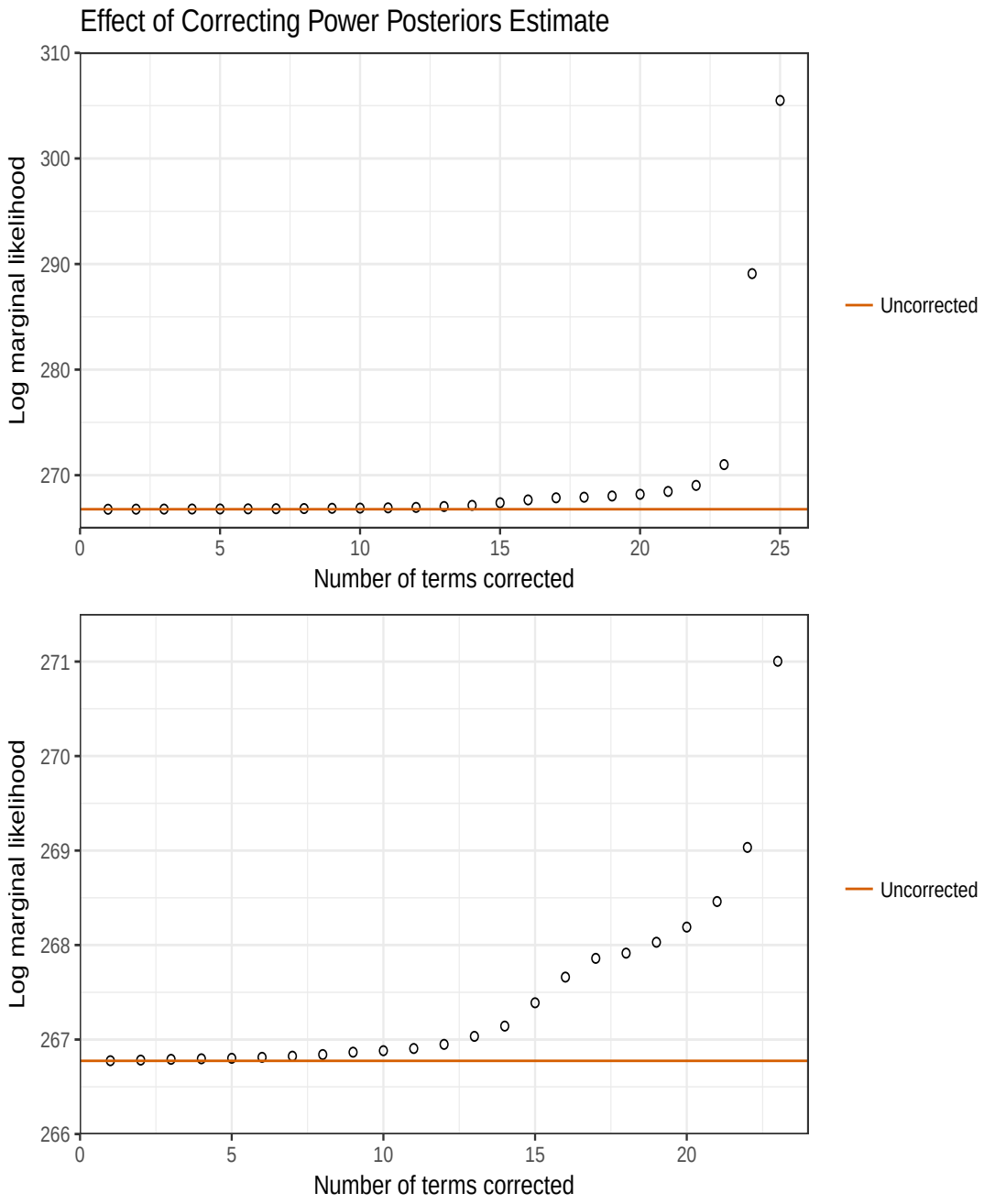
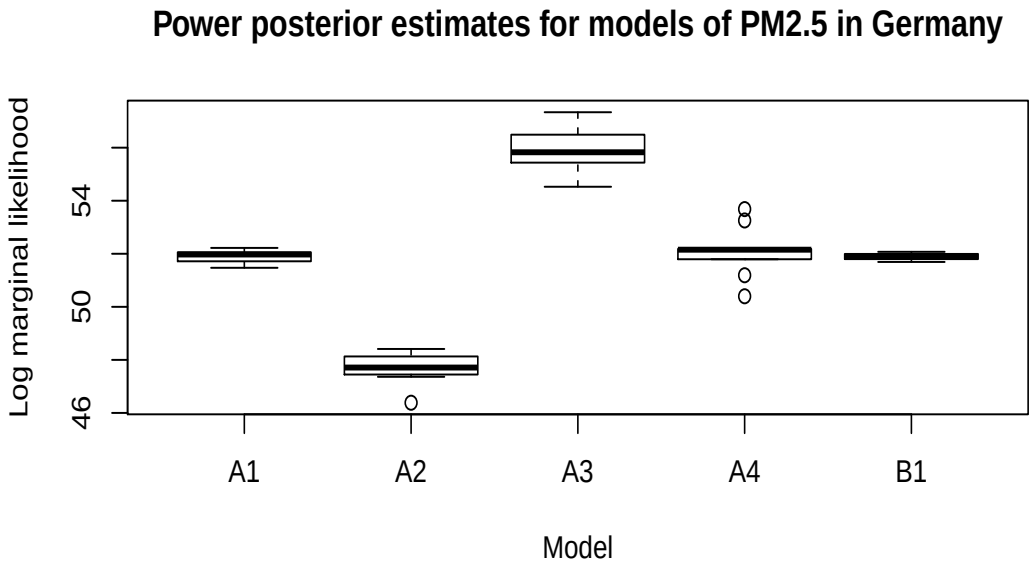



Figure 4-4: These graphs highlight an anomaly with the power posteriors correction for model A4 with data from China. The graphs are of one estimate, showing the cumulative effect of successive corrections. The curve is expected to resemble that of Figure 2-6. The magnitudes of the final five or six corrections are far larger than they ought to be.



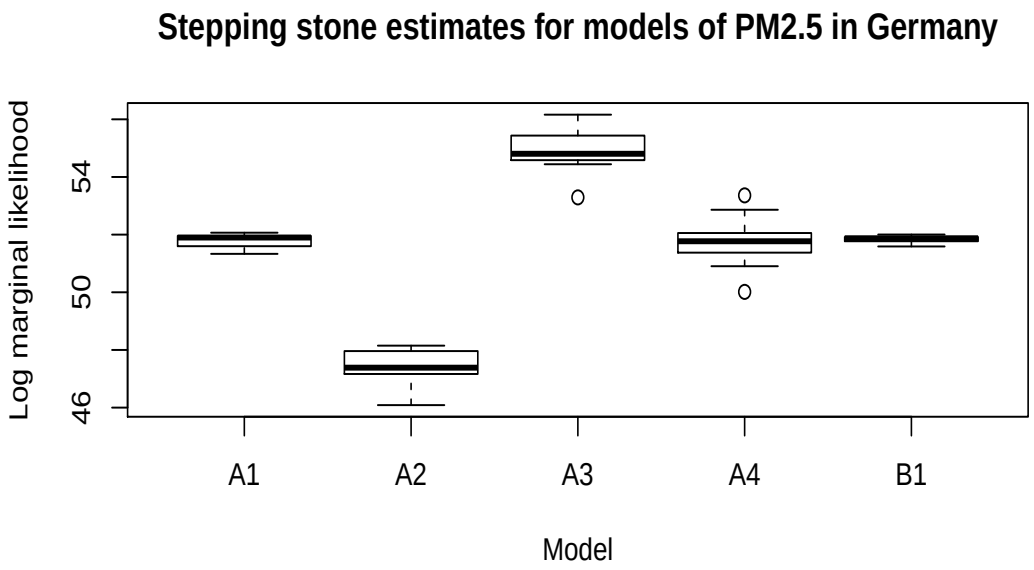




Figure 4-5: The results of the two tempering methods for the data from Germany. Ten estimates were made with each method.

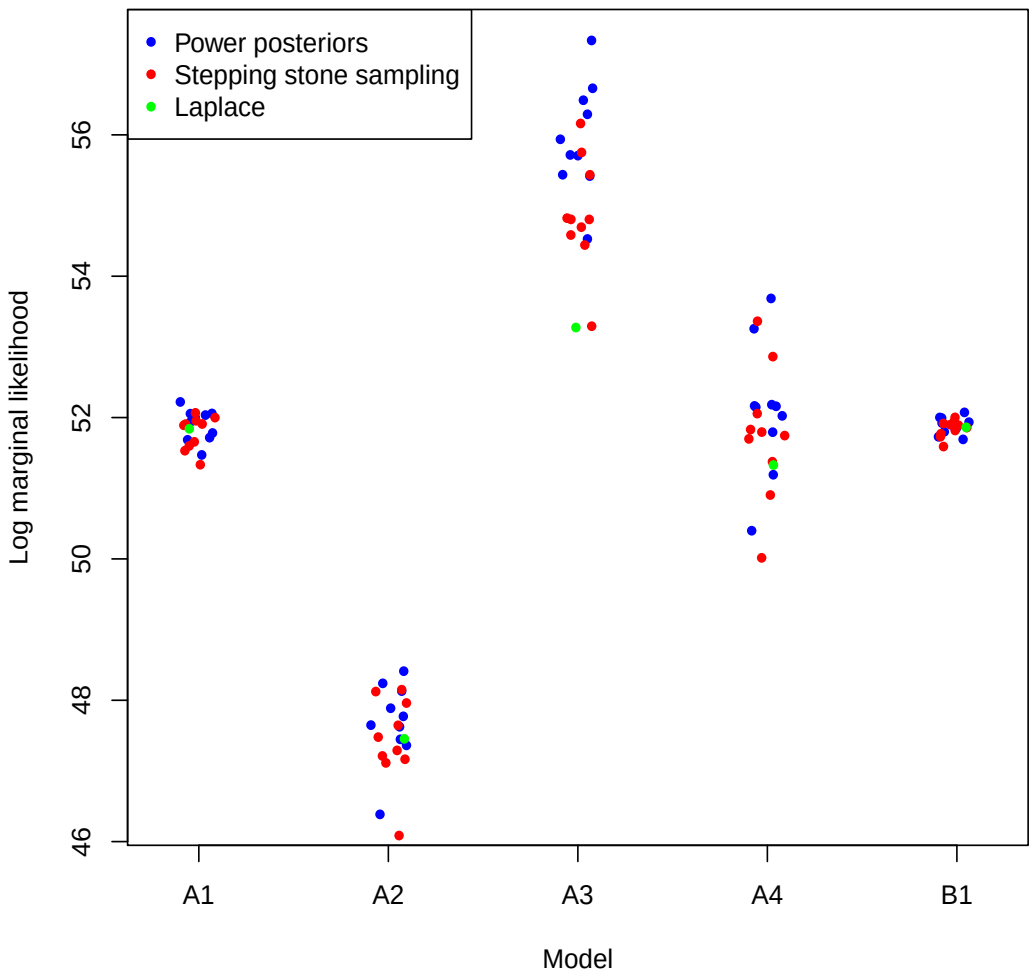
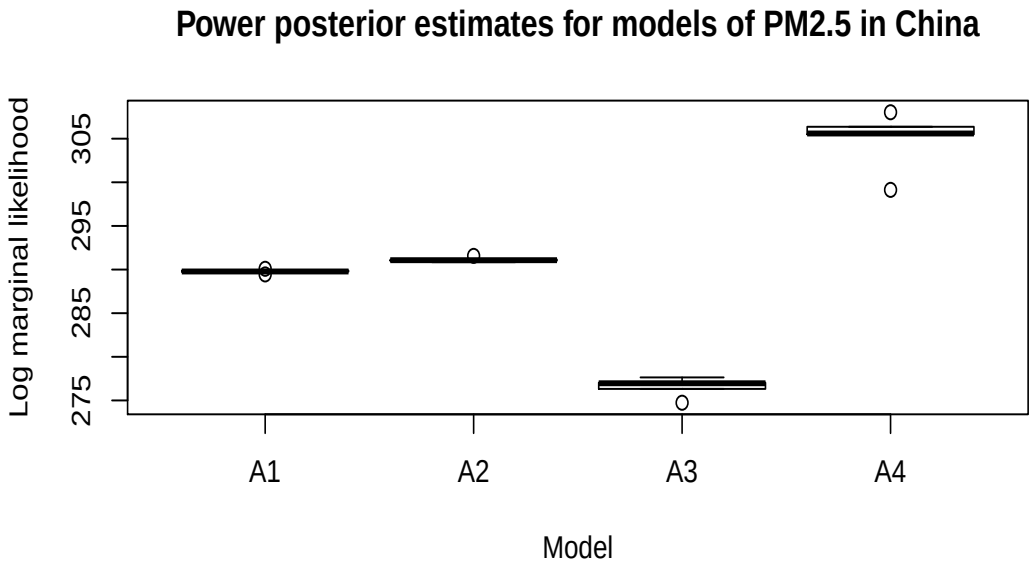



Figure 4-6: Side by side comparison of estimates of the log marginal likelihood from all three methods for the Germany pollution data.



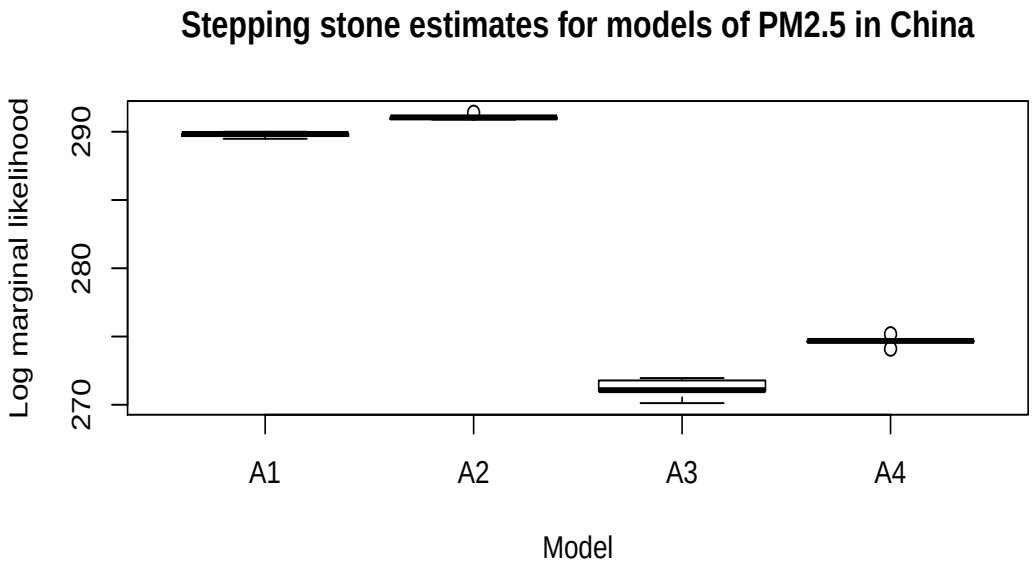




Figure 4-7: The results of the two tempering methods for the data from China. Ten estimates were made with each method.

# Chapter 5

# Summary

In chapter two the methods for estimating the marginal likelihood were reviewed. Use of the methods for estimating the marginal likelihood was examined for two examples, from the initial simple example, the four most promising methods were selected to test on a second example with real data. In this example the Laplace method, stepping-stone sampling and the method of power posteriors performed best. These methods would then be the methods taken forward to use for estimating the marginal likelihood for spatial models in the preceding chapters. Additionally there was a subsection on the effect of improving power posteriors via a correction term, where it was found that it improved the results of this method.

In Chapter three some concepts and definitions of spatial statistics were introduced, chiefly Gaussian Random Fields which are the principal tools used to model continuous spatial data. Gaussian random fields are defined by their mean and covariance functions, the Matérn class of covariance functions is the most widely used. Of particular interest is the exponential covariance model that is often used to model pollution so the focus was on models with this structure. An efficient approach for calculating the densities required for the Markov Chain Monte Carlo simulation was also detailed in this chapter. The number of temperatures and size of the MCMC samples to use for the tempering methods was investigated and it was found that increasing the number of temperatures beyond fifty gave no noticeable improvement in precision. Increasing from twenty to forty thousand samples at each temperature reduced the standard error of the estimates, more testing would need to be done to confirm the size of the effect, and to test intermediate sample sizes. Implication of the tempering methods was tested on data simulated from a Gaussian random field model with powered exponential covariance function; performance was similar for the two methods

except for the case $k = 1.5$ where the variance was higher for power posteriors. The power posteriors also gave consistently lower values for the estimate of the marginal likelihood relative to steppingstone sampling.

In chapter four a model for real spatial data was examined; the data concerned concentrations of $PM_{2.5}$ in Germany and China. A Gaussian random field model was constructed with exponential covariance function. Estimates for the marginal likelihood were obtained from the Laplace method, steppingstone sampling and power posteriors. Laplace method and steppingstone sampling estimated the log marginal likelihood consistently, the method of power posteriors gave inconsistent results which may be due to problems with the MCMC sampling used. The Laplace method was much faster than the two tempering methods so could be recommended from this exploratory case study.

To conclude, the Laplace method was the best performing method for calculating the marginal likelihood for spatial models of the Matérn class. Given the straightforward implementation and low computation time of the Laplace method, anyone looking at fitting Gaussian random field models could easily calculate accurate marginal likelihoods for each model that they would want to compare. Of the tempering methods, steppingstone sampling is recommended. However the extra computation time of the tempering methods did not result in noticeably improved accuracy of the estimates relative to the Laplace method. There will be cases where the Laplace method will not be an effective method for estimating the marginal likelihood however, and steppingstone sampling could be the best method in these cases.

For further work it would be interesting to see the efficacy of the methods for estimating the marginal likelihood of spatio-temporal models or hierarchical models. Comparison of model selection through Bayes factors to model selection by other model selection criterion such as BIC would be another area of investigation.

# Appendix A

# Code

## A.1 Power posteriors and steppingstone sampling for an exponential model

The following function, `dmvnrm_arma`, calculates the density of a multivariate Gaussian distribution given the mean and standard deviation.

```
#include <RcppArmadillo.h>

const double log2pi = std::log(2.0 * M_PI);

// [[Rcpp::depends("RcppArmadillo")]]
// [[Rcpp::export]]
double dmvnrm_arma(arma::rowvec x,
                       arma::rowvec mean,
                       arma::mat sigma,
                       bool logd = false) {

  int xdim = x.size();
  double out;
  arma::mat rooti = arma::trans(arma::inv(trimatu(arma::chol(sigma))));
  double rootisum = arma::sum(log(rooti.diag()));
  double constants = -(static_cast<double>(xdim)/2.0) * log2pi;

    arma::vec z = rooti * arma::trans( x - mean) ;
    out      = constants - 0.5 * arma::sum(z%z) + rootisum;

  if (logd == false) {
    out = exp(out);
  }
  return(out);
}
```

Listing A.1: $dmvnrm\_arma.cpp$

The function `edmvnrm_arma` calculates the density of a multivariate Gaussian distribution given the Eigen decomposition of the covariance matrix with two scaling values $v$ and $\sigma$, and the mean. The covariance matrix is of the form $\Sigma_{ij} = \sigma exp(-\phi$

```
#include <RcppArmadillo.h>
using namespace Rcpp;
using namespace arma;

const double log2pi = std::log(2.0 * datum::pi);

// [[Rcpp::depends("RcppArmadillo")]]
// [[Rcpp::export]]
arma::mat edmvn_arma(arma::rowvec y,
                     arma::rowvec Mu,
                     arma::rowvec lam,
                     arma::mat U,
                     double v,
                     double sigma) {

  int ydim = y.size();
  arma::mat out;
  arma::rowvec D =  pow(sigma * lam + v,-1);
  arma::rowvec x = y - Mu;
  arma::mat s = x * U * (trans(D)%(trans(U)*trans(x)));
  double ldet = sum(log(sigma*lam+v));

  out = (-0.5*ydim*log2pi -0.5*ldet - 0.5*s);
  return(out);
}
```

Listing A.2: $eigendmvnorm.cpp$

Code for the Metropolis Hastings algorithm used in Section 4.1 for generating samples at each temperature for model A1. A slightly modified function was used for each model.

```

sourceCpp(file="eigendmvnorm.cpp")

WHO_MH3X1 <- function(init,t=1,n,data,jump,lphi,ED,priorv,priorm,coeffs){
  m=length(init)
  subcount <- rep(0,m)
  subratios <- matrix(0,m,n)
  subproposed <- matrix(0,m,n)
```

```
  subout <- matrix(init,m,n)
  lls <- rep(0,n)

  tjump=rep(0,m)
  tjump[1] = min(jump[1]*(t^-0.5),3) #Beta 0
  for(i in 2:3){tjump[i] = min(jump[i]*(t^-0.5),1)} #variance terms
  for(i in 4:m){tjump[i] = min(jump[i]*(t^-0.5),3)} #all other Beta coefficients

  mvtmu = coeffs %*% c(init[1],init[4:m]) # initial mean, X%*%Beta

  oldldev =  edmvn_arma(Mu=mvtmu,lam=ED$values,U=ED$vectors,v=exp(init[3]),sigma
      =exp(init[2]),y=data)
  llold = t*oldldev + dnorm(init[1],priorm[1],priorv[1],log=T) +
    dnorm(init[2],priorm[2],priorv[2],log=T) + dnorm(init[3],priorm[3],priorv
        [3],log=T) +
    dnorm(init[4],priorm[4],priorv[4],log=T) + dnorm(init[5],priorm[5],priorv
        [5],log=T)
  lls[1] = oldldev

  for(i in 2:n){
    ##Beta0
    subout[1,i] = subout[1,i-1] + rnorm(1,0,tjump[1])

    mvtmu = coeffs %*% c(subout[1,i],subout[4:m,i-1]) # mean, X%*%Beta

    newldev = edmvn_arma(Mu=mvtmu,lam=ED$values,U=ED$vectors,v=exp(subout[3,i
        -1]),sigma=exp(subout[2,i-1]),y=data)

    llnew = t*newldev + dnorm(subout[1,i],priorm[1],priorv[1],log=T) +
      dnorm(subout[2,i-1],priorm[2],priorv[2],log=T) + dnorm(subout[3,i-1],
          priorm[3],priorv[3],log=T) +
      dnorm(subout[4,i-1],priorm[4],priorv[4],log=T) + dnorm(subout[5,i-1],
          priorm[5],priorv[5],log=T)

    if(exp(llnew-llold) > runif(1)){
      subcount[1] = subcount[1]+1
      llold = llnew
      oldldev=newldev
    }else{
      subout[1,i] = subout[1,i-1]
    }

    ##Beta1,2,...,m
    for(j in 4:m){
      subout[j,i] = subout[j,i-1] + rnorm(1,0,tjump[j])

      if(j != m){
        mvtmu = coeffs %*% c(subout[1,i],subout[4:j,i],subout[(j+1):m,i-1]) #
            initial mean, X%*%Beta
      }else{mvtmu = coeffs %*% c(subout[1,i],subout[4:j,i])}
      ### takes beta0, beta1...j latest value (ith) which have been updated,
```

```
          betaj+1...m value from last loop (i-1) as haven't been updated yet
53
54
55      newldev = edmvn_arma(Mu=mvtmu,lam=ED$values,U=ED$vectors,v=exp(subout[3,i
            -1]),sigma=exp(subout[2,i-1]),y=data)
56
57      llnew = t*newldev + dnorm(subout[1,i],priorm[1],priorv[1],log=T) +
58        dnorm(subout[2,i-1],priorm[2],priorv[2],log=T) + dnorm(subout[3,i-1],
              priorm[3],priorv[3],log=T) +
59        dnorm(subout[4,i],priorm[4],priorv[4],log=T) +   c(j<5, j>=5) %*% rbind(
               dnorm(subout[5,i-1],priorm[5],priorv[5],log=T), dnorm(subout[5,i],
              priorm[5],priorv[5],log=T))
60
61    #     if(pmax[1] < llnew){pmax=c(llnew,subout[1,i],subout[2,i-1],subout[3,i
          -1],subout[4,i],subout[5,i+j-5])}
62
63        if(exp(llnew-llold) > runif(1)){
64        subcount[j] = subcount[j]+1
65        llold = llnew
66        oldldev=newldev
67      }else{
68        subout[j,i] = subout[j,i-1]
69      }
70    }
71
72    ##Sigma
73    subout[2,i] = subout[2,i-1] + rnorm(1,0,tjump[2])
74
75    mvtmu = coeffs %*% c(subout[1,i],subout[4:m,i]) # initial mean, X%*%Beta
76
77    newldev = edmvn_arma(Mu=mvtmu,lam=ED$values,U=ED$vectors,v=exp(subout[3,i
          -1]),sigma=exp(subout[2,i]),y=data)
78
79    llnew = t*newldev + dnorm(subout[1,i],priorm[1],priorv[1],log=T) +
80      dnorm(subout[2,i],priorm[2],priorv[2],log=T) + dnorm(subout[3,i-1],priorm
            [3],priorv[3],log=T) +
81      dnorm(subout[4,i],priorm[4],priorv[4],log=T) + dnorm(subout[5,i],priorm
            [5],priorv[5],log=T)
82
83    if(exp(llnew-llold) > runif(1)){
84      subcount[2] = subcount[2]+1
85      llold = llnew
86      oldldev=newldev
87      #if(subout[2,i] > 4){print(c(subout[2,i-1],subout[2,i]))}
88    }else{
89      subout[2,i] = subout[2,i-1]
90    }
91    ##V
92    subout[3,i] = subout[3,i-1] + rnorm(1,0,tjump[3])
93
94    mvtmu = coeffs %*% c(subout[1,i],subout[4:m,i]) # initial mean, X%*%Beta
```

```
95
96      newldev = edmvn_arma(Mu=mvtmu,lam=ED$values,U=ED$vectors,v=exp(subout[3,i]),
            sigma=exp(subout[2,i]),y=data)
97
98      llnew = t*newldev + dnorm(subout[1,i],priorm[1],priorv[1],log=T) +
99        dnorm(subout[2,i],priorm[2],priorv[2],log=T) + dnorm(subout[3,i],priorm
              [3],priorv[3],log=T) +
100       dnorm(subout[4,i],priorm[4],priorv[4],log=T) + dnorm(subout[5,i],priorm
              [5],priorv[5],log=T)
101     if(exp(llnew-llold) > runif(1)){
102       subcount[3] = subcount[3]+1
103       llold = llnew
104       oldldev=newldev
105     }else{
106       subout[3,i] = subout[3,i-1]
107     }
108
109     lls[i]=oldldev
110   }
111   Ell = mean(lls)
112   #print(Ell)
113   return(list(subsamples=subout,Ej=Ell,likelihoods=lls,submoves=subcount))
114 }
```

The following function calculates the power posterior and stepping stone sampling estimates of the marginal likelihood. Requires previous function which samples at each temperature.

```
1  WHOgrfppss3 = function(n = 1000, vj=c(2,0.15,0.15),d.mat,coeffs,
2                         thet=seq(from=0.1,to=6,by=0.005),kappa=1,
3                         N=1,thedata,priorv=c(10,2,2),priorm=c(0,0,0),temps=seq
                              (1,0,by=-0.05)^5,FUN=WHO_MH3X1){
4
5   ed = eigen(exp(-(d.mat*thet)^kappa))
6   ev = ed$values
7   U = ed$vectors
8
9   m = length(priorm)
10
11  burn = n/10
12  M=length(temps)
13  count = array(0,dim=c(N,M,m))
14  likelihoods=array(0,dim=c(N,M,n-burn))
15  out = array(c(rep(0,(n-burn)*N*M),rep(0,(n-burn)*N*M),rep(1,(n-burn)*N*M)),dim
        = c(N,M,n-burn,m))
16  ppterms <- matrix(0,N,M-1)
17  ppcorrection <- matrix(0,N,M-1)
18  ppmarginal = rep(0,N)
```

```
  ssterms = matrix(0,N,M-1)
  ssmarginal = rep(0,N)
  l <- dim(d.mat)[1]
  Ejs <- rep(0,M)
  vars <-matrix(0,N,M)

  for(i in 1:N){

    initialvalues = rep(0,m)

    #get MCMC samples#
    for(j in 1:M){
      samples <- FUN(init=initialvalues,n=n,data=thedata,jump = vj,t=temps[j],
          lphi=length(thet),ED=ed,priorv=priorv,priorm=priorm,coeffs=coeffs)
      count[i,j,] <- samples$submoves
      out[i,j,,] <- t(samples$subsamples[,(burn+1):n])
      likelihoods[i,j,] <- samples$likelihoods[(burn+1):n]
      initialvalues = samples$subsamples[,n]
      Ejs[j] = samples$Ej
      vars[i,j] = var(samples$likelihoods[(burn+1):n])
    #  postmax = samples$thetahat
      # plot(temps[1:j],Ejs[1:j],type="b")

      if(j>1){
        ppterms[i,j-1] <- (temps[j-1]-temps[j])*(Ejs[j-1]+Ejs[j])*0.5
        ppcorrection[i,j-1] = (vars[i,j]-vars[i,j-1])*((temps[j]-temps[j-1])^2)/
            12

        logmax = max(likelihoods[i,j-1,])
        ssterms[i,j-1] = (temps[j-1]-temps[j])*logmax + log((1/(n-burn))*sum(exp
            ((temps[j-1]-temps[j])*(likelihoods[i,j,]-logmax))))
      }
    }
    ppmarginal[i] = sum(ppterms[i,]) + sum(ppcorrection[i,])
    ssmarginal[i] = sum(ssterms[i,])
  }
  return(list(marginalpp=ppmarginal,marginalss=ssmarginal,points=out,moves=count
      ,ssterm=ssterms,ppterm=ppterms,ppcorrection=ppcorrection,likelihoods=
      likelihoods,variances = vars))
}
```

# Bibliography

[Abbey et al., 1995] Abbey, D. E., Ostro, B. E., Petersen, F., and Burchette, R. J. (1995). Chronic respiratory symptoms associated with estimated long-term ambient concentrations of fine particulates less than 2.5 microns in aerodynamic diameter (pm2. 5) and other air pollutants. *Journal of exposure analysis and environmental epidemiology*, 5(2):137–159.

[Abrahamsen, 1997] Abrahamsen, P. (1997). *A review of Gaussian random fields and correlation functions*. Norsk Regnesentral/Norwegian Computing Center.

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle,[w:] proceedings of the 2nd international symposium on information, bn petrow, f. *Czaki, Akademiai Kiado, Budapest*.

[Aukema et al., 2008] Aukema, B. H., Carroll, A. L., Zheng, Y., Zhu, J., Raffa, K. F., Dan Moore, R., Stahl, K., and Taylor, S. W. (2008). Movement of outbreak populations of mountain pine beetle: influences of spatiotemporal patterns and climate. *Ecography*, 31(3):348–358.

[Betancourt, 2017] Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

[Boldo et al., 2006] Boldo, E., Medina, S., Le Tertre, A., Hurley, F., Mücke, H.-G., Ballester, F., Aguilera, I., et al. (2006). Apheis: Health impact assessment of long-term exposure to PM2.5 in 23 european cities. *European Journal of Epidemiology*, 21(6):449–458.

[Burnham and Anderson, 2004] Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2):261–304.

[Calderhead and Girolami, 2009] Calderhead, B. and Girolami, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics & Data Analysis*, 53(12):4028–4045.

[Cambardella and Karlen, 1999] Cambardella, C. and Karlen, D. (1999). Spatial analysis of soil fertility parameters. *Precision Agriculture*, 1(1):5–14.

[Chatfield, 2016] Chatfield, C. (2016). *The analysis of time series: an introduction*. CRC press.

[Chib, 1995] Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321.

[Chib and Jeliazkov, 2001] Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453):270–281.

[Chopin and Robert, 2010] Chopin, N. and Robert, C. P. (2010). Properties of nested sampling. *Biometrika Advance Access*, pages 1–15, doi: 10.1093/biomet/asq021.

[Cressie, 1991] Cressie, N. (1991). *Statistics for spatial data*. John Wiley & Sons.

[Diggle et al., 1998] Diggle, P. J., Tawn, J., and Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350.

[Etz and Wagenmakers, 2015] Etz, A. and Wagenmakers, E.-J. (2015). Origin of the Bayes factor. *arXiv preprint arXiv:1511.08180*.

[Fann et al., 2012] Fann, N., Lamson, A. D., Anenberg, S. C., Wesson, K., Risley, D., and Hubbell, B. J. (2012). Estimating the national public health burden associated with exposure to ambient PM2.5 and ozone. *Risk Analysis*, 32(1):81–95.

[Feroz and Hobson, 2008] Feroz, F. and Hobson, M. (2008). Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, 384(2):449–463.

[Finkenst adt et al., 2006] Finkenst adt, B., Held, L., and Isham, V. (2006). *Statistical methods for spatio-temporal systems*. CRC Press.

[Finley et al., 2007] Finley, A. O., Banerjee, S., and Carlin, B. P. (2007). spbayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4):1.

[Friel et al., 2014] Friel, N., Hurn, M., and Wyse, J. (2014). Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24(5):709–723.

[Friel and Pettitt, 2008] Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607.

[Friel and Wyse, 2012] Friel, N. and Wyse, J. (2012). Estimating the evidence - a review. *Statistica Neerlandica*, 66(3):288–308.

[Gorsich and Genton, 2000] Gorsich, D. J. and Genton, M. G. (2000). Variogram model selection via nonparametric derivative estimation. *Mathematical Geology*, 32(3):249–270.

[Gräler et al., 2016] Gräler, B., Pebesma, E., and Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *The R Journal*, 8:204–218.

[Hoeting et al., 1999] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, pages 382–401.

[Hug et al., 2016] Hug, S., Schwarzfischer, M., Hasenauer, J., Marr, C., and Theis, F. J. (2016). An adaptive scheduling scheme for calculating Bayes factors with thermodynamic integration using Simpson's rule. *Statistics and Computing*, pages 1–15.

[Huijnen et al., 2010] Huijnen, V., Williams, J., Weele, M. v., Noije, T. v., Krol, M., Dentener, F., Segers, A., Houweling, S., Peters, W., Laat, J. d., et al. (2010). The global chemistry transport model TM5: description and evaluation of the tropospheric chemistry version 3.0. *Geoscientific Model Development*, 3(2):445–473.

[Jagla, 2010] Jagla, E. (2010). Realistic spatial and temporal earthquake distributions in a modified Olami-Feder-Christensen model. *Physical Review E*, 81(4):046117.

[Kacenelenbogen et al., 2006] Kacenelenbogen, M., Léon, J.-F., Chiapello, I., and Tanré, D. (2006). Characterization of aerosol pollution events in France using ground-based and polder-2 satellite data. *Atmospheric Chemistry and Physics*, 6(12):4843–4849.

[Kass and Raftery, 1995] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

[Kuha, 2004] Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological methods & research*, 33(2):188–229.

[Lartillot and Philippe, 2006] Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207.

[Lindgren et al., 2010] Lindgren, F., Lindström, J., and Rue, H. (2010). *An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach*. Mathematical Statistics, Centre for Mathematical Sciences, Faculty of Engineering, Lund University.

[Liu et al., 2005] Liu, Y., Sarnat, J. A., Kilaru, V., Jacob, D. J., and Koutrakis, P. (2005). Estimating ground-level PM2.5 in the eastern United States using satellite remote sensing. *Environmental Science & Technology*, 39(9):3269–3278.

[López-Granados et al., 2002] López-Granados, F., Jurado-Expósito, M., Atenciano, S., García-Ferrer, A., de la Orden, M. S., and García-Torres, L. (2002). Spatial variability of agricultural soil parameters in southern Spain. *Plant and Soil*, 246(1):97–105.

[Marinari and Parisi, 1992] Marinari, E. and Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, 19(6):451.

[Mukherjee et al., 2006] Mukherjee, P., Parkinson, D., and Liddle, A. R. (2006). A nested sampling algorithm for cosmological model selection. *The Astrophysical Journal Letters*, 638(2):L51.

[Neal, 2001] Neal, R. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.

[Neal, 1996] Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366.

[Newton and Raftery, 1994] Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48.

[Ostro and Chestnut, 1998] Ostro, B. and Chestnut, L. (1998). Assessing the health benefits of reducing particulate matter air pollution in the United States. *Environmental Research*, 76(2):94–106.

[Podur et al., 2003] Podur, J., Martell, D. L., and Csillag, F. (2003). Spatial patterns of lightning-caused forest fires in Ontario, 1976–1998. *Ecological Modelling*, 164(1):1–20.

[Raftery, 1986] Raftery, A. E. (1986). A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society, Series B*, 48(2):249–250.

[Raftery et al., 1996] Raftery, A. E. et al. (1996). Hypothesis testing and model selection via posterior simulation. *Markov chain Monte Carlo in practice*, pages 163–188.

[Robert and Wraith, 2009] Robert, C. P. and Wraith, D. (2009). Computational methods for Bayesian model choice. In *American Institute of Physics Conference Series*, volume 1193, pages 251–262.

[Rue and Held, 2005] Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

[Rue et al., 2009] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

[Sacks et al., 1989] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science*, pages 409–423.

[Sain and Cressie, 2007] Sain, S. R. and Cressie, N. (2007). A spatial model for multivariate lattice data. *Journal of Econometrics*, 140(1):226–259.

[Schabenberger and Gotway, 2004a] Schabenberger, O. and Gotway, C. A. (2004a). *Statistical methods for spatial data analysis*. CRC press.

[Schabenberger and Gotway, 2004b] Schabenberger, O. and Gotway, C. A. (2004b). *Statistical methods for spatial data analysis*. CRC Press.

[Schwarz et al., 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Shaddick et al., 2018] Shaddick, G., Thomas, M. L., Green, A., Brauer, M., Donkelaar, A., Burnett, R., Chang, H. H., Cohen, A., Dingenen, R. V., Dora, C., et al. (2018). Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(1):231–253.

[Shaddick and Wakefield, 2002] Shaddick, G. and Wakefield, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(3):351–372.

[Shaddick and Zidek, 2014] Shaddick, G. and Zidek, J. V. (2014). A case study in preferential sampling: Long term monitoring of air pollution in the uk. *Spatial Statistics*, 9:51–65.

[Shaddick and Zidek, 2015] Shaddick, G. and Zidek, J. V. (2015). *Spatio-Temporal Methods in Environmental Epidemiology*. CRC Press.

[Shaw et al., 2007] Shaw, R., Bridges, M., and Hobson, M. (2007). Clustered nested sampling: efficient Bayesian inference for cosmology. *Monthly Notices of the Royal Astronomical Society*, 378.

[Skilling et al., 2006] Skilling, J. et al. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–859.

[Smith et al., 1988] Smith, J. W., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. (1988). Using the ADAP learning algorithm to forecast the onset of Diabetes Mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association.

[Spiegelhalter et al., 2014] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):485–493.

[Spiegelhalter et al., 2002] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

[Tierney and Kadane, 1986] Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86.

[Tobler, 1970] Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(sup1):234–240.

[Van Dingenen et al., 2014] Van Dingenen, R., Leitao, J., and Dentener, F. (2014). A multi-metric global source-receptor model for integrated impact assessment of climate and air quality policy scenarios. In *EGU General Assembly Conference Abstracts*, volume 16.

[Van Donkelaar et al., 2016] Van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., Lyapustin, A., Sayer, A. M., and Winker, D. M. (2016). Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. *Environmental science & technology*, 50(7):3762–3772.

[Van Wagtendonk and Cayan, 2008] Van Wagtendonk, J. W. and Cayan, D. R. (2008). Temporal and spatial distribution of lightning strikes in California in relation to large-scale weather patterns. *Fire Ecology*.

[Wang and Anderson, 2011] Wang, Y. and Anderson, K. R. (2011). An evaluation of spatial and temporal patterns of lightning-and human-caused forest fires in Alberta, Canada, 1980–2007. *International Journal of Wildland Fire*, 19(8):1059–1072.

[Wasserman, 2000] Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107.

[Watkins, 2004] Watkins, D. S. (2004). *Fundamentals of matrix computations*, volume 64. John Wiley & Sons.

[WHO et al., 2006] WHO et al. (2006). World Health Organisation air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment.

[Worm et al., 2005] Worm, B., Sandow, M., Oschlies, A., Lotze, H. K., and Myers, R. A. (2005). Global patterns of predator diversity in the open oceans. *Science*, 309(5739):1365–1369.

[Xie et al., 2011] Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160.

[Zhu et al., 2010] Zhu, J., Huang, H.-C., and Reyes, P. E. (2010). On selection of spatial linear models for lattice data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):389–402.